

Adversarial vulnerability of machine-learning surrogates for computational fluid dynamics

Jian-Nan Chen^{1,2,*}, Xiao-Yan Cao^{3,*}, and Jun-Jie Zhang^{2,§}

¹*School of Electronic Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China*

²*Northwest Institute of Nuclear Technology, No. 28 Pingyu Road, Xi'an 710024, Shaanxi, China*

³*School of Nuclear Engineering, Rocket Force University of Engineering, No. 2 Tongxin Road, Xi'an 710025, Shaanxi, China*



(Received 23 December 2025; accepted 26 May 2026; published 1 July 2026)

Machine-learning surrogates are increasingly used to accelerate computational fluid dynamics (CFD), yet their robustness to small, structured input perturbations remains insufficiently characterized. Conventional adversarial training is largely impractical for artificial intelligence for CFD (AI4CFD), since perturbed inputs alter the underlying partial differential equation (PDE) solutions and require costly resimulation for valid labels. We propose a systematic *post hoc* robustness audit and apply it to the Fourier neural operator (FNO) and UNet on three canonical two-dimensional flows—laminar tube flow, transient dam break, and cylinder wake—using Gaussian, random, and fast gradient sign method (FGSM) perturbations together with normalized mean-squared error, attack gain, and signed-bias diagnostics. Adversarial vulnerability is found to arise from a coupling between flow physics and architectural inductive bias: The transient dam break exhibits heavy-tailed gains with rare amplification of order 10^2 , while steadier flows remain bounded. A spectral analysis with a proposed attack separability index shows that FNO behaves like a spectral solver with global mode coupling, whereas UNet acts as a dissipative low-pass scheme, with the distinction modulated by PDE type. A physical-realizability audit, including a divergence-free Helmholtz projection control, rules out continuity violation and distributional shift, indicating that the vulnerability is intrinsic to the learned operator rather than a generalization artifact. We therefore argue that robustness auditing should accompany accuracy reporting in AI4CFD, and provide a practical three-step audit protocol.

DOI: [10.1103/n5kp-jjcp](https://doi.org/10.1103/n5kp-jjcp)

I. INTRODUCTION

The application of artificial intelligence (AI), particularly deep learning, has gained considerable attention in computational fluid dynamics (CFD) in recent years [1–6]. AI models, particularly deep neural networks, are being explored as surrogates for traditional numerical solvers, enabling approximate solutions to the complex partial differential equations (PDEs) that govern fluid behavior [7]. These models offer the potential to accelerate simulations, particularly in cases involving high-dimensional or complex flow phenomena, where traditional solvers become computationally prohibitive [1–4,8,9].

Additionally, machine learning-based operator learning frameworks, such as the deep operator network (DeepONet) [10] and Fourier neural operator (FNO) [11,12], learn mappings that generalize across entire function spaces, enabling AI models to adapt to a broader range of flow conditions. Recent studies have also extended these approaches toward

physically consistent operator learning and differentiable field reconstruction [8,13], showing promise for integrating AI within scientific and engineering workflows such as optimization, control, and uncertainty quantification [2,4,6,14].

Despite these advances, modern AI models are susceptible to small, structured perturbations in their inputs, which can lead to unpredictable errors in the output [15–20] (e.g., the fast gradient sign method, FGSM). This vulnerability presents a challenge in AI for CFD (AI4CFD), where even minor deviations in input data—such as changes in boundary conditions, material properties, or geometry—can result in large errors in the predictions of flow behavior.

Unlike traditional numerical solvers, which tend to exhibit stable and predictable behavior under small perturbations, AI models may amplify such disturbances due to the complex, nonlinear nature of partial differential equations (PDEs). This vulnerability can be particularly concerning in real-world applications. However, it remains unclear how this fragility is shaped by the interplay between the inherent instability of the flow physics and the inductive bias of the neural architecture, necessitating a study that disentangles these factors.

In image classification, adversarial training can leverage labeled perturbations to enhance robustness [16,18,21]. However, a central challenge in AI4CFD is that input perturbations fundamentally alter the solutions to the governing PDEs. Generating corresponding ground-truth labels for these perturbed systems requires costly high-fidelity simulations or experiments [22,23], which are computationally prohibitive at scale.

*These authors contributed equally to this work.

†Contact author: chen1994@alumni.sjtu.edu.cn

‡Contact author: mosescao@mail.ustc.edu.cn

§Contact author: zjacob@mail.ustc.edu.cn

Consequently, AI4CFD lacks readily available ground truth for adversarially perturbed inputs. This fundamental constraint renders label-dependent adversarial training strategies largely impractical for most CFD applications.

This paper aims to propose a systematic robustness audit for AI-driven CFD surrogates. Specifically, we evaluate two representative AI architectures—FNO and UNet [24]—which are trained autoregressively on three distinct flow scenarios: a developing laminar tube flow, a transient dam-break flow, and flow around a circular cylinder from the CFD bench [25]. We focus on these two architectures because they provide high-accuracy clean rollouts in the benchmark while representing two structurally distinct inductive biases: global spectral coupling in FNO and localized convolutional processing in UNet. Other operator-learning architectures, including DeepONet, are discussed in the context of this bias spectrum in Sec. III E. We apply three perturbation schemes, including Gaussian noise, random perturbations, and gradient-aligned adversarial attacks [16–18], to evaluate the robustness of the models.

Our results reveal that adversarial vulnerability is not a universal property but arises from an interaction between the flow’s physical characteristics and the network’s architectural bias. Specifically, dam-break flows exhibit heavy-tailed gain distributions with rare but extreme relative amplification factors of 100–200, whereas tube and cylinder flows show more moderate degradation. These findings demonstrate that robustness cannot be inferred from accuracy alone. We therefore argue that robustness auditing and reporting should become a standard practice in evaluating CFD surrogates, particularly for safety-critical engineering applications.

II. RESULTS

We evaluate how machine-learning surrogates for CFD respond to small, structured input perturbations. FNO and UNet are trained autoregressively to predict the velocity field at the next time step according to the current flow state and scenario parameters. We test three canonical flows: a laminar tube flow with varying fluid properties, a transient dam break with gravity-driven interface collapse, and a cylinder wake flow covering both steady and vortex-shedding regimes. For each scenario, geometry and boundary conditions are held fixed while fluid properties vary within specified ranges; see Sec. V for details.

Model responses are evaluated over multistep rollouts using three complementary diagnostics: the normalized mean-squared error (NMSE), an attack gain quantifying the relative specificity of gradient-aligned FGSM perturbations compared with clean and equal-budget untargeted baselines, and a signed bias quantifying whether parameter or flow perturbations dominate. The overall workflow is illustrated in Fig. 1. Together these experiments provide a physics-aware stress test revealing how flow physics and model architecture jointly influence the robustness of AI surrogates for CFD.

A. Qualitative rollouts and illustrative case studies

To gain an intuitive understanding of how the FGSM attack affects fluid-model predictions, we use the dam-break scenario as a representative example. Figure 2 illustrates

representative rollouts for this case. Each row corresponds to a different type of input perturbation—clean, FGSM perturbed, random noise, and Gaussian noise—applied with the same perturbation budget ϵ . The left column shows the initial (clean/perturbed) inputs at $t = 0$, the middle column presents the corresponding model predictions at step = 3, and the right column displays the ground-truth reference. Although the initial conditions appear nearly identical, the FGSM-perturbed rollout exhibits large divergence, revealing failure modes consistent with the vulnerabilities discussed above.

Similar deviations occur in the cylinder and tube scenarios. The characteristics of these errors are scenario dependent: Dam break exhibits pronounced interface/topology differences and localized high-error regions; cylinder primarily shows phase shifts and asymmetry in the wake; and tube displays a gradual mean-bias drift. Note that the predictions correspond to recognizable, physically meaningful deviations rather than mere numerical noise. Throughout the following, perturbations are applied by default to the initial input of all models using a small deviation.

B. Global vulnerability patterns via gain distributions

We begin by characterizing the overall vulnerability landscape of the two surrogate architectures across the three benchmark flows. For each model-scenario pair, we compute the attack gain [Eq. (10)], which measures the relative specificity of gradient-aligned FGSM perturbations compared with clean and equal-budget untargeted baselines. Figure 3 summarizes the distributions of gain for all six combinations, pooling values over all evaluated rollout steps and perturbation budgets.

Two qualitative patterns are revealed. First, the dam-break case displays a heavy-tailed vulnerability distribution: While the majority of samples show modest amplification, a small fraction of cases produce very large gains (on the order of 10^2). Because gain is a relative metric, these outliers should not be interpreted from the ratio alone. We therefore complement the gain distributions with absolute NMSE- ϵ curves and a gain-versus-absolute-error diagnostic in Figs. 12 and 14. These analyses show that the high-gain tail in dam break includes configurations with substantial FGSM-induced NMSE, supporting their interpretation as rare but severe failures in strongly transient flows. Second, both the tube and cylinder scenarios exhibit tighter gain distributions: Most gain values are concentrated below ~ 30 , with rapidly decaying tails. These cases should therefore be interpreted primarily as more bounded relative adversarial sensitivity unless they are also accompanied by large absolute NMSE.

A further observation is that the gain distributions share similar qualitative shapes between the two architectures for each flow scenario: Both FNO and UNet exhibit comparable concentration, spread, and tail behavior. This implies that the underlying flow physics—rather than the architecture alone—governs the susceptibility to rare, large-amplitude failures. Specifically, the strongly transient, nonlinear dam-break flow leads to heavy-tailed risk, while the steady tube flow and the cylinder flow—which spans both steady and periodic vortex-shedding regimes—constrain perturbation amplification to a more bounded range.

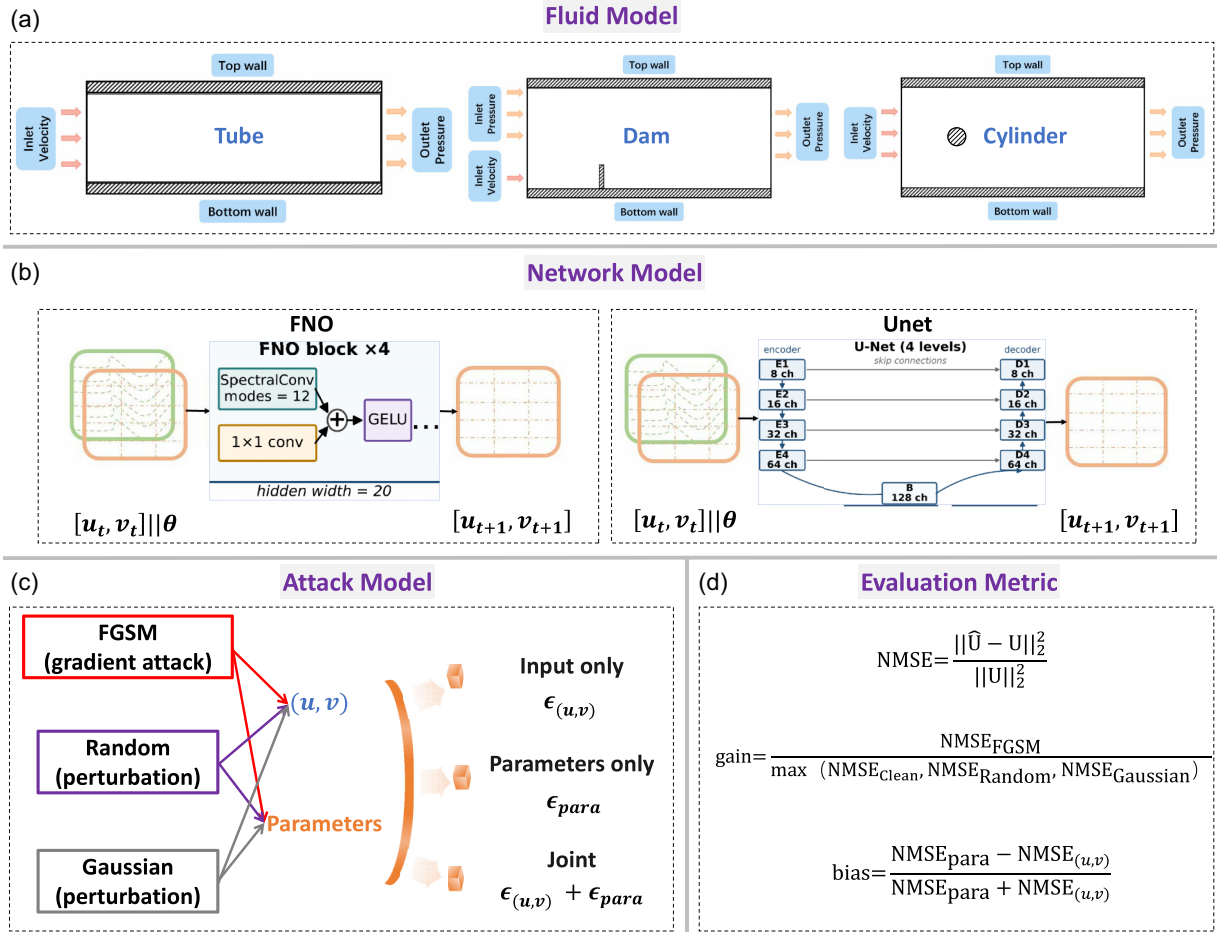


FIG. 1. Robustness assessment framework for neural networks. (a) Physical settings for three canonical CFD problems: tube flow, dam break, and flow past a cylinder. (b) Neural-network architectures: FNO and UNet. The orange grids represent velocity fields (u, v) , while green grids denote parameter channels. Both surrogates are trained to predict $[\hat{u}_{t+1}, \hat{v}_{t+1}]$ from the input $[u_t, v_t] \parallel \theta$. During evaluation, only the initial step is perturbed; predictions are then fed back autoregressively to simulate multistep rollouts. (c) Attack configurations: Gaussian (gray frame) and random (purple frame) noise as baselines, and FGSM (red) as the adversarial method, all under a shared perturbation budget ϵ . Perturbation targets include parameter-only, flow-state-only, and joint parameter-flow attacks. (d) Evaluation metrics: normalized mean squared error (NMSE), attack gain (FGSM relative to clean and equal-budget untargeted baselines), and signed bias (NMSE difference between parameter and flow targets).

These conclusions are conditional on the specific parameter ranges studied. For the tube flow, the Reynolds number range $\text{Re} = \rho u_b d / \mu$ spans approximately 0.1–90, placing all simulations firmly in the laminar regime. The dam-break cases cover a broad range of density ratios (up to 1000:1) and Reynolds numbers from 1 to 10 000, encompassing both laminar and turbulent regimes. Cylinder flows operate in the Reynolds number range of 10–500, covering both steady and unsteady wake regimes. The observed robustness patterns should be interpreted within these specific flow regimes.

C. Channel sensitivity: parameters versus flow

Next, we study which parts of the input tensor are most effective as attack targets: (1) scalar scenario parameters (geometry, inlet speed, viscosity, etc.), (2) the instantaneous flow channels (u, v) , or (3) joint perturbations that affect both. Figure 4 presents grouped boxplots of **gain** for these three target choices, plotted separately for each scenario and architecture.

The box plots reveal clear, scenario-dependent patterns:

(1) *Cylinder*. For both FNO and UNet, perturbing the instantaneous velocity field (u, v) produces larger gains than perturbing parameters. This is consistent with the cylinder wake flow exhibiting high sensitivity to velocity perturbations [26]: In regimes with vortex shedding ($\text{Re} > 47$), small changes in the velocity field can shift shedding phase and symmetry [27,28]; even in steady regimes ($\text{Re} < 47$), perturbations may disrupt the wake development [29,30]. In both cases, instantaneous flow changes tend to produce larger downstream errors than gradual parametric shifts.

(2) *Dam break*. The two architectures behave differently. FNO is highly susceptible to (u, v) perturbations, while parameter-only attacks have little effect; in contrast, UNet shows more comparable sensitivity to parameters and to (u, v) . This divergence likely reflects both representation differences and the inherent physics of free-surface flows [31,32]. FNO's global spectral coupling can amplify coherent field perturbations that disrupt the evolving free surface—A

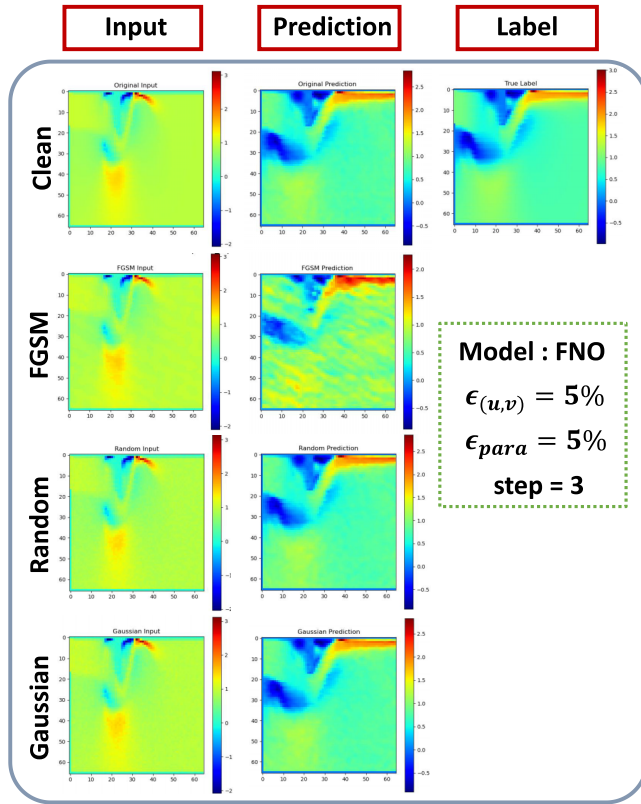


FIG. 2. Representative clean vs perturbed rollouts for dam-break model (similar input gives different predictions). Columns show (left) four input variants—clean, FGSM perturbed, random noise, and Gaussian noise; (center) the corresponding predictions with step = 3; and (right) the ground-truth solution. Despite nearly identical initial conditions at $t = 0$, the FGSM-perturbed results diverge, exhibiting scenario-specific failure modes consistent with the vulnerabilities documented above.

phenomenon well known in dam-break simulations where small initial disturbances can lead to significantly different interface evolution. UNet’s local convolutional features [33] moderate that global coupling due to their spatial locality, while allowing parameter channels to exert a relatively larger influence on the bulk flow behavior [34].

(3) *Tube*. For the laminar tube flow—a steady, fully developed flow—the velocity profile depends analytically on parameters like the Reynolds number. The observed high parameter sensitivity of UNet therefore serves as a validation of the proposed bias metric, confirming its ability to capture the expected physical influence. Interestingly, FNO exhibits the opposite tendency, remaining more sensitive to (u, v) -field perturbations. This divergence highlights the role of architectural bias: The near-steady, diffusive physics makes sustained parameter shifts particularly consequential for architectures that process information locally [33], whereas models with global spectral coupling [12] retain a higher sensitivity to coherent structures in the velocity field, even in this regime. Thus, while the baseline parameter sensitivity is expected, the contrasting response of the two models reveals their distinct inductive biases.

The target-channel trends are further examined using absolute NMSE- ϵ curves grouped by attack target (Fig. 13).

These curves show that the gain-based channel sensitivity is not solely a normalization artifact. In dam break and cylinder, flow-field perturbations generally produce larger absolute FGSM errors than parameter-only perturbations, whereas tube exhibits weaker adversarial specificity despite non-negligible absolute NMSE at large perturbation budgets.

These findings demonstrate that the most critical perturbation channel depends on both the flow regime and the model’s inductive bias. Therefore, robustness audits must assess parameter and flow-field channels, rather than relying on a single test target.

D. Signed bias between attack channels

We quantify the relative importance of parameter versus flow-field attacks using the signed bias metric [Eq. (11)]. Figure 5(a) presents the signed bias distributions for each model-scenario pair, while Fig. 5(b) shows the paired relationship between signed bias and channel severity for the same matched parameter-only and flow-field-only attacks.

The distributions show three succinct patterns. First, for the dam break and the cylinder wake flows the bias distributions are predominantly negative for dam-FNO and cylinder-UNet/FNO, confirming that perturbations applied to the instantaneous velocity field (u, v) typically produce larger normalized errors than parameter-only perturbations. Second, dam-UNet displays a broad, near-symmetric bias distribution. This should not be interpreted as evidence of robustness; rather, it indicates that parameter-only and flow-field-only attacks produce comparable NMSE values, with the dominant channel varying across samples. Third, tube-UNet is distinct in showing a clear positive bias, i.e., parameter perturbations tend to dominate for that architecture and regime.

The signed bias should be interpreted as a directionality metric, not as a severity metric. Values near $+1$ or -1 identify a dominant attack channel, whereas values near zero indicate balanced channel sensitivity. However, a near-zero bias can correspond either to balanced robustness or to balanced vulnerability, depending on the absolute NMSE scale. We therefore pair the signed bias with the channel severity,

$$\text{severity}_{\text{ch}} = \frac{1}{2}(\text{NMSE}_{\text{para}} + \text{NMSE}_{(u,v)}),$$

computed from the same matched parameter-only and flow-field-only attacks at the same perturbation budget and rollout step. The resulting bias-severity view in Fig. 5(b) shows the absolute error scale associated with each bias value.

Dam-UNet provides a useful example of this distinction. Its bias distribution is broad and approximately centered around zero, indicating that parameter-only and flow-field-only attacks often produce comparable NMSE values rather than a single consistently dominant channel. In the paired bias-severity view, these balanced cases mostly occupy a moderate severity range. Thus, dam-UNet is best characterized by balanced channel sensitivity with moderate absolute error, rather than by a single dominant failure channel.

E. Temporal compounding of errors across autoregressive rollouts

Autoregressive inference can amplify a single initial perturbation through repeated feedback. Figure 6 shows the

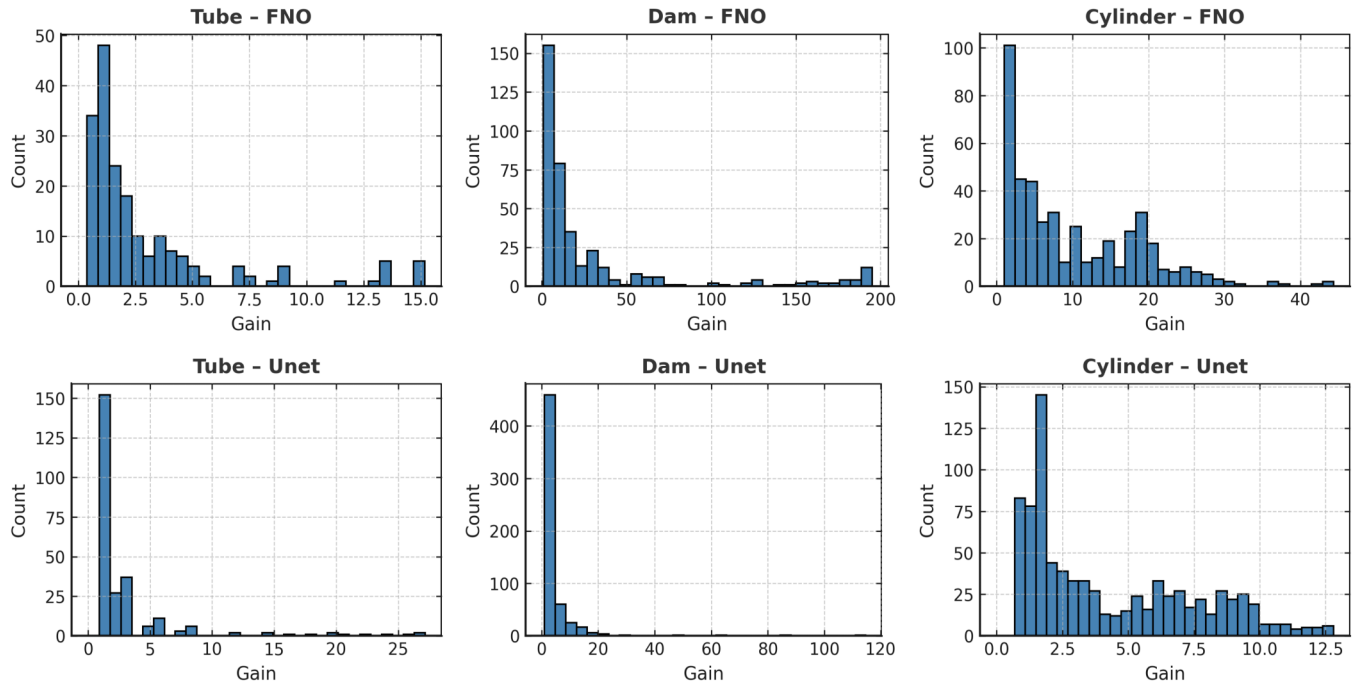


FIG. 3. Distributions of attack **gain** for six model-scenario pairs. Top row: FNO on (left to right) tube, dam break, cylinder. Bottom row: UNet on the same three scenarios. Each panel pools **gain** values over all evaluated rollout steps and perturbation budgets. Gain is computed as the FGSM NMSE relative to the larger of the clean, random, and Gaussian NMSE baselines under the same perturbation budget. Key observations: (i) dam break exhibits heavy-tailed behavior with occasional extreme relative amplifications (rare events reaching **gain** ~ 100 – 200); (ii) tube and cylinder show substantially tighter distributions with most gains concentrated below ~ 30 ; (iii) within each scenario the FNO and UNet share the same qualitative distribution shape. Because gain is a relative metric, the corresponding absolute NMSE magnitudes are reported in Figs. 12 and 14.

results for the dam-break case of NMSE versus autoregressive step for several values of ϵ ; qualitatively similar behavior roughly holds for the other scenarios.

Two features stand out. First, NMSE grows with both prediction step and attack strength: Larger ϵ produces faster early growth and higher long-horizon error, and even small first-step perturbations can compound into large multistep deviations. We note that the $\epsilon(u, v) = 0.01$ curve (blue dashed)

nearly coincides with the clean baseline (black) in both panels: At this magnitude, the perturbation-induced error is comparable to the model’s intrinsic single-step prediction error, so the perturbed and unperturbed trajectories are statistically indistinguishable. Second, the architectures differ in the rate and form of this compounding. The FNO typically exhibits a rapid rise in NMSE within the first few steps for moderate-to-large ϵ , indicating fast coherent propagation of gradient-aligned

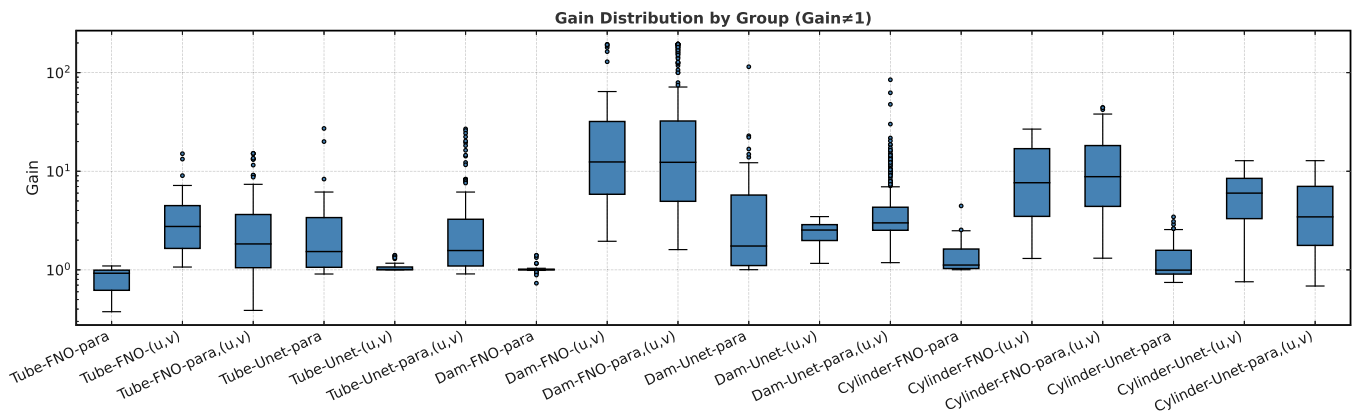


FIG. 4. Box plots of attack **gain** grouped by attack target. The box plot groups **gain** distributions by attack target—parameter-only, (u, v) -only, and joint perturbations—for each model-scenario pair, pooling values over all evaluated rollout steps and perturbation budgets. Key observations: In the cylinder case, (u, v) attacks dominate; in the dam-break case FNO is strongly dominated by (u, v) attacks while UNet is more balanced; in the tube case UNet is more sensitive to parameter perturbations whereas FNO shows larger sensitivity to (u, v) . The corresponding absolute NMSE- ϵ curves grouped by target are provided in Fig. 13.

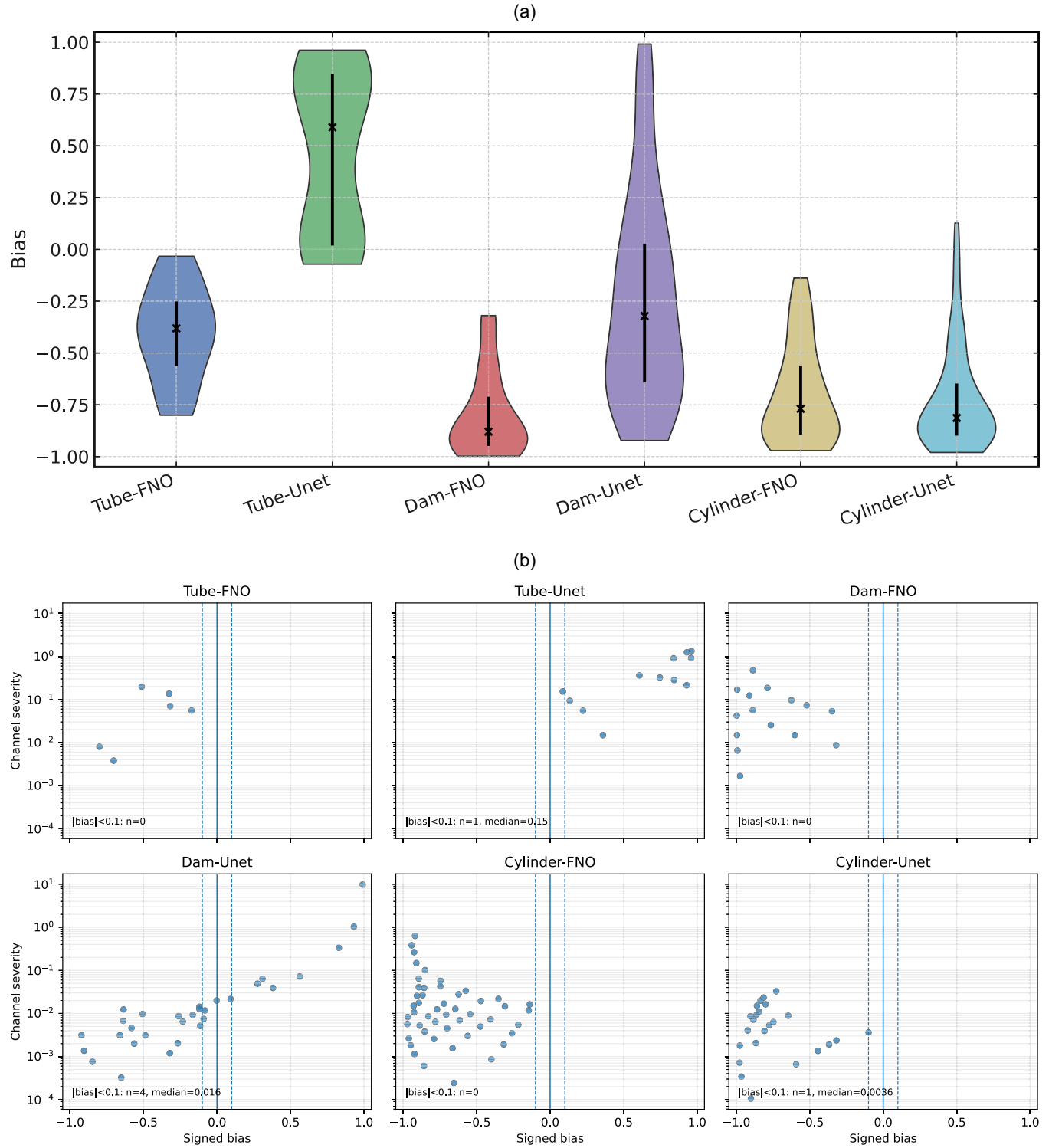


FIG. 5. Signed channel bias and paired channel severity under parameter-only and flow-field-only attacks. (a) Signed bias distributions for each model-scenario pair, $\text{bias} = (\text{NMSE} \times \text{para} - \text{NMSE} \times (u, v)) / (\text{NMSE} \times \text{para} + \text{NMSE} \times (u, v))$, pooled over all rollout steps and perturbation budgets. Negative values indicate flow-field-dominated degradation, whereas positive values indicate parameter-dominated degradation; white bars denote medians and interquartile ranges. (b) Paired bias-severity diagnostic from the same matched FGSM attacks. Each point represents one perturbation budget and rollout step, with signed bias on the horizontal axis and channel severity, $\text{severity} \times \text{ch} = \frac{1}{2}(\text{NMSE} \times \text{para} + \text{NMSE}_{(u,v)})$, on the vertical axis. The shaded region marks near-balanced cases, $|\text{bias}| \leq 0.1$. This representation distinguishes balanced robustness, near-zero bias with low severity, from balanced vulnerability, near-zero bias with high severity. Together with Figs. 4 and 13, these diagnostics separate attack-channel directionality from absolute error magnitude.

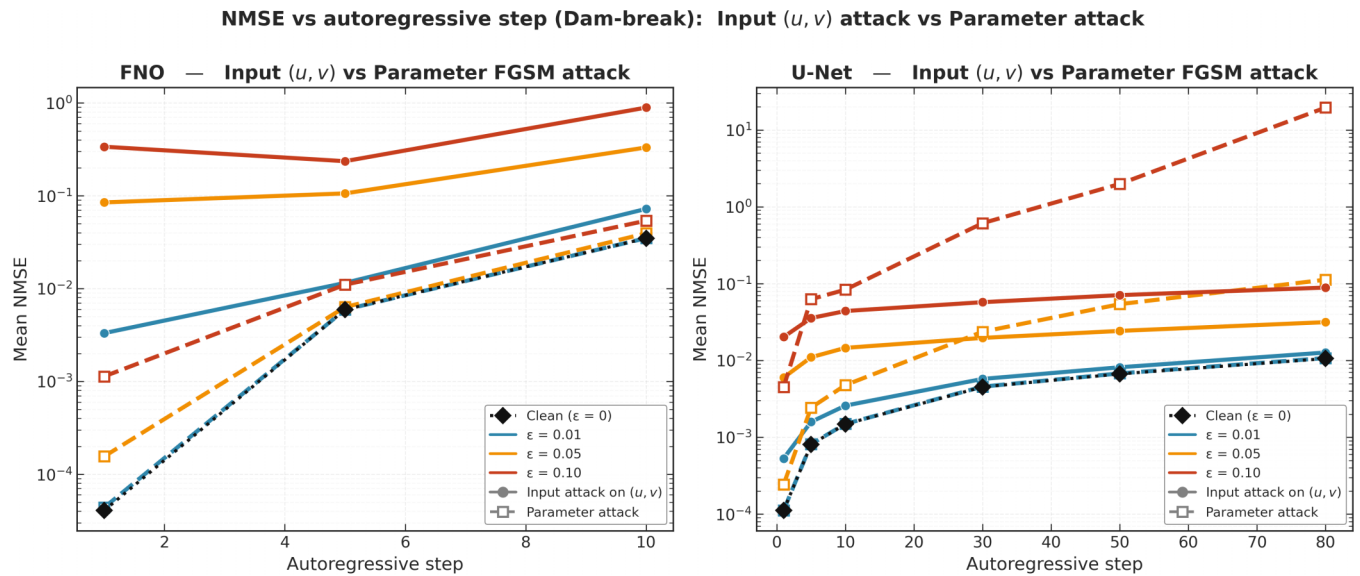


FIG. 6. NMSE vs autoregressive step for multiple FGSM budgets (dam break). Each panel plots mean NMSE (y axis) against autoregressive prediction step (x axis) for a set of FGSM budgets ϵ (different coloured lines). Left: FNO; right: U-Net. NMSE increases with both step and attack strength. FNO typically shows sharp early escalation for moderate-to-large ϵ , while U-Net tends to accumulate error more gradually and often attains a plateau at longer horizons.

disturbances through its global spectral layers. After 10 steps, the predictions of FNO—including those from the clean input—become distorted significantly, and thus the comparison is omitted after that step. The U-Net generally accumulates error more gradually and often reaches a plateau at longer horizons, suggesting localized error diffusion and partial attenuation of perturbation strength. The plateau does not imply correctness, but indicates a statistically steady error amplitude under these conditions rather than unbounded divergence.

The observed temporal compounding patterns are indicative of the architectures' inherent numerical properties [3]. The rapid error escalation in FNO mirrors the behavior of a spectral solver [35], where perturbation energy is transferred globally across scales, leading to coherent error growth that mimics physical instabilities. Consequently, once the error is introduced, it propagates instantaneously across the domain. Conversely, U-Net's convolutional and pooling layers [24] introduce an effect analogous to numerical dissipation or artificial viscosity in traditional CFD schemes [36]. This inherent damping tends to smear out small-scale errors, preventing coherent growth and leading to a statistical plateau where the perturbation input is balanced by the model's diffusive characteristics.

For practical assessment we therefore recommend reporting NMSE curves across a small grid of attack budgets (as in Fig. 6), since single-step metrics alone do not capture long-horizon stability.

III. DISCUSSION

Table I summarizes the main behavioral patterns observed in our experiments. Our findings indicate that the fragility of learned surrogates is not random; rather, it follows predictable patterns governed by the coupling of physical dynamics and neural numerical properties. Specifically, two

main conclusions can be drawn. First, the underlying *flow physics* plays a key role in determining the upper bound of vulnerability: Highly transient and nonlinear dynamics (e.g., dam break) lead to heavy-tailed failure risks, whereas flows that remain steady or relax to a regular wake (e.g., tube, cylinder) tend to limit error amplification to a narrower range. Second, the *model architecture* dictates how these perturbations evolve spatially: FNO's global spectral coupling allows errors to propagate coherently across the domain and enhances cross-channel interactions, while U-Net's locally focused features tend to diffuse perturbations, resulting in slower accumulation.

A. Mechanistic interpretation: numerical inductive bias

The distinct responses of FNO and U-Net can be understood by mapping their architectural features to classical numerical methods in CFD. To move beyond qualitative analogy, we quantitatively verify this mapping through spectral analysis of the error fields under adversarial and baseline perturbations (Figs. 7–9 and Table II).

FNO as a spectral method. The FNO operates via global Fourier layers, which mathematically resemble spectral methods used in turbulence simulations. While efficient, spectral methods are notoriously sensitive to discontinuities and high-frequency perturbations. Our spectral analysis confirms that FNO inherits this property: Its global receptive field allows local perturbations to propagate coherently across the entire domain via mode coupling. Quantitatively, FNO maintains a nontrivial ASI across the *entire* wavenumber range—including high κ (e.g., $\text{ASI} \approx 5.84$ at the high-frequency band for dam break; Table II)—demonstrating that adversarial energy is redistributed across all spectral modes rather than being attenuated at any particular scale. This is the signature of global mode coupling: Fourier convolutions produce an

TABLE I. Concise summary of robustness patterns across scenarios and surrogate architectures. Results are conditional on the geometric and nondimensional parameter ranges reported in Sec. V. Note: Cylinder flow aggregates both steady and unsteady wake regimes based on the parameter ranges.

Scenario	Model	Dominant attack channel	Relative gain/tail behavior	Temporal compounding
Tube (laminar, diffusive)	FNO	(u, v) slightly more damaging (coherent spectral response)	Moderate gains; concentrated tails (most $\text{gain} < 20$)	Gradual NMSE growth; modest compounding.
Tube (laminar, diffusive)	UNet	Parameters dominate (operating-point bias)	Moderate gains; tight tails	Slow NMSE growth; errors frequently saturate to a bias plateau.
Dam break (strongly transient)	FNO	(u, v) strongly dominant	Heavy-tailed vulnerability: rare extreme events ($\text{gain} \sim 10^2$)	Rapid early NMSE escalation for moderate ϵ .
Dam break (strongly transient)	UNet	Mixed sensitivity (sample dependent; both channels relevant)	Heavy tails but fewer extreme outliers than FNO	NMSE grows and often reaches a statistical plateau.
Cylinder (wake flow)	FNO	(u, v) dominant (wake sensitivity)	Tightly concentrated to moderate tails (most $\text{gain} < 20$)	Early-stage NMSE rise; FNO shows super-additive coupling as (u, v) budget increases.
Cylinder (wake flow)	UNet	(u, v) dominant (wake development)	Similar concentrated gains as FNO, generally less extreme outliers	Slower compounding than FNO.

architecture-characteristic error footprint that is largely independent of the input perturbation type, as evidenced by the near-collapse of all four attack spectra in the raw error spectra (Fig. 11).

UNet as a dissipative scheme. In contrast, the UNet relies on localized convolution kernels and pooling operations. From a numerical analysis perspective, these operations act analogously to finite difference stencils with added artificial viscosity or numerical dissipation. Our error-spectrum analysis reveals two quantitative signatures of this dissipative behavior:

(1) *Amplitude attenuation:* The total error energy of UNet is $3\text{--}12\times$ lower than that of FNO across all attack types and flow scenarios (Fig. 8), consistent with uniform amplitude damping across frequencies.

(2) *Selective sensitivity to perturbation structure:* UNet exhibits sharp ASI decay with wavenumber κ , collapsing toward $\text{ASI} \approx 1$ (no separation from baseline noise) at high wavenumbers (Fig. 7). This confirms that convolutional and pooling layers act as a low-pass amplitude filter; isotropic noise (Gaussian/random) is effectively suppressed, but gradient-aligned adversarial perturbations—which are inherently low frequency and structurally coherent—bypass the dissipative mechanism and produce disproportionate damage at low κ .

Importantly, this selective sensitivity means that UNet does not exhibit uniform robustness to all perturbation types. Rather, its dissipative bias provides effective protection against isotropic, high-frequency disturbances (energy amplification ratio $\text{EAR} \approx 1.5\text{--}3\times$ for Gaussian noise) while remaining vulnerable to structured, gradient-aligned attacks that concentrate energy in the low-frequency modes where dissipation is weakest ($\text{EAR} \approx 11\text{--}38\times$ for FGSM). This distinction—between mode-selective attenuation and uniform robustness—is critical for practical deployment and underscores that numerical dissipation, while stabilizing, is not a complete defense against adversarial threats.

B. Role of flow physics

The vulnerability is further modulated by the mathematical nature of the governing equations. Our spectral analysis provides the first quantitative verification of this modulation through the ASI gap—defined as the difference in mean ASI between FNO and UNet—which serves as a scalar measure of how strongly the architectural distinction manifests under each flow regime.

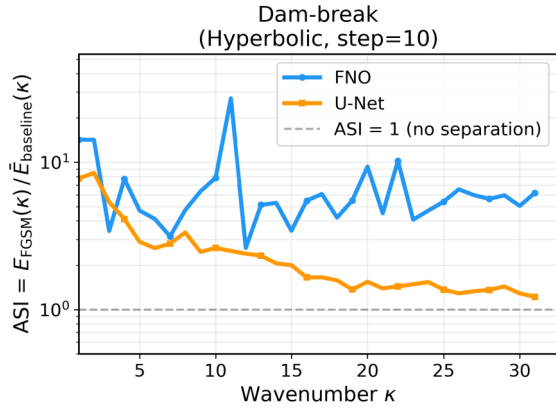
The ASI gap exhibits a strict ordering with PDE type (Fig. 9):

(1) *Hyperbolic (dam break):* $\text{ASI gap} = 4.27$. The free-surface flow is dominated by advection and characterized by positive Lyapunov exponents [37], meaning any surrogate faces a steep challenge in maintaining trajectory stability. The large ASI gap confirms that FNO’s spectral coupling maximally amplifies adversarial structure in this regime, while UNet’s dissipation provides partial (but incomplete) protection.

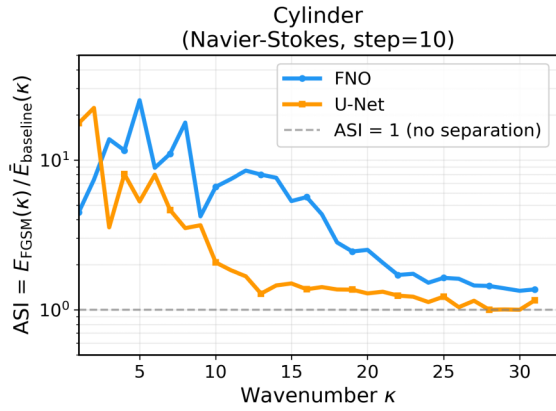
(2) *Mixed (cylinder):* $\text{ASI gap} = 2.48$. The Navier-Stokes flow combines advective and diffusive mechanisms, producing an intermediate regime where the architectural distinction is present but attenuated relative to the hyperbolic case.

(3) *Parabolic (tube):* $\text{ASI gap} = 0.08 \approx 0$. The diffusion-dominated flow naturally attenuates all disturbances over time, providing a physical buffer that renders the architectural choice nearly irrelevant. Both FNO and UNet converge to essentially identical spectral behavior (Fig. 11), confirming that when the PDE itself supplies sufficient dissipation, the network’s inductive bias becomes secondary.

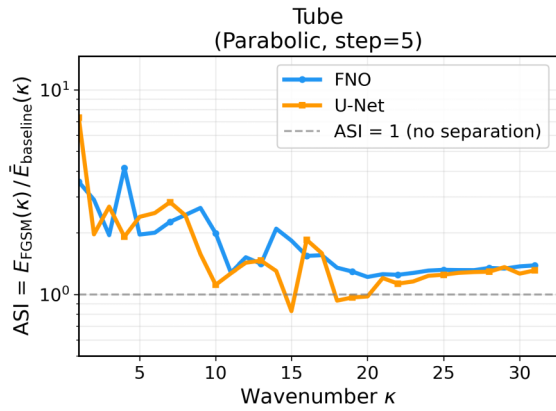
This ordering constitutes a nontrivial prediction of the mechanistic framework: UNet’s implicit dissipation provides a structural advantage *only* when the underlying PDE does not already supply sufficient diffusion. The parabolic tube case serves as a critical falsification test—If the ASI gap had remained large even in a diffusion-dominated regime, the



(a) Dam-break (step=10)



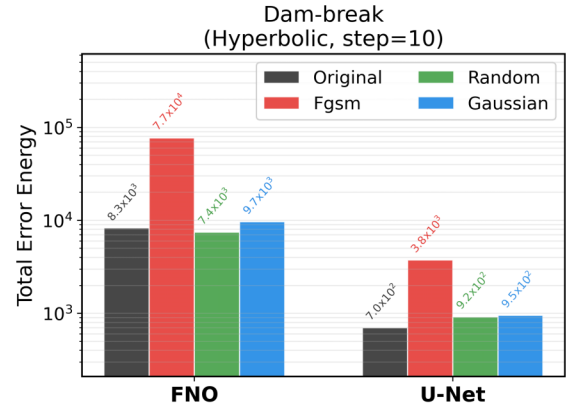
(b) Cylinder (step=10)



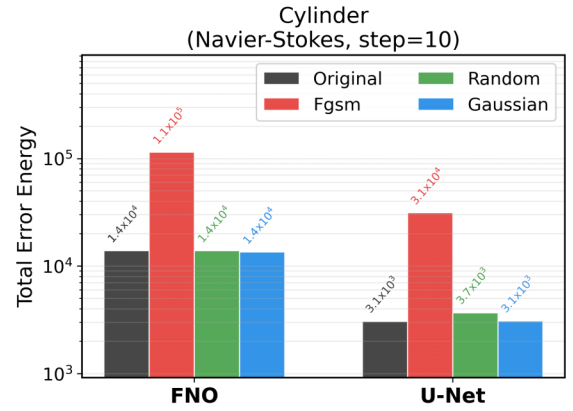
(c) Tube (step=5)

FIG. 7. Attack Separability Index (ASI) for FNO and UNet across wavenumbers κ . ASI is defined as $E_{\text{FGSM}}(\kappa) / \bar{E}_{\text{baseline}}(\kappa)$, where the baseline averages clean, random, and Gaussian error spectra. $\text{ASI} = 1$ (dashed line) indicates no separation from random noise. FNO maintains elevated ASI across all κ (global mode coupling), while UNet's ASI collapses to 1 at high κ (dissipative low-pass filtering). The intermodel gap decreases from dam break to tube, tracking the transition from hyperbolic to parabolic dynamics.

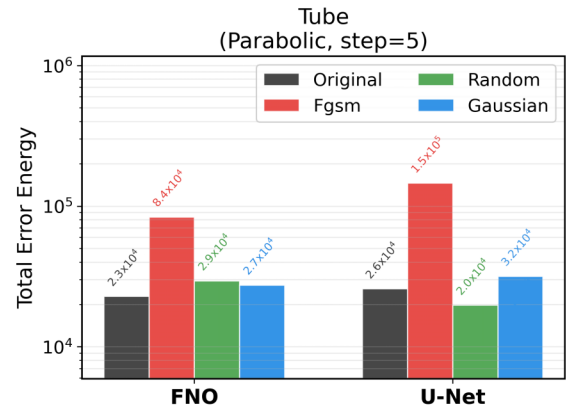
“UNet-as-dissipation” analogy would have been invalidated. Its near-vanishing value instead provides strong confirmatory evidence.



(a) Dam-break (step=10)



(b) Cylinder (step=10)



(c) Tube (step=5)

FIG. 8. Total error energy (integrated power spectral density, log scale) for each model-attack combination. UNet exhibits 3–12 $\times$ lower total energy than FNO in dam break and cylinder, consistent with its dissipative amplitude attenuation. In tube (parabolic), both models converge to comparable energy levels, confirming that PDE-intrinsic diffusion renders the architectural distinction negligible.

C. Implications for safety-critical applications

Current AI4CFD paradigms often prioritize accuracy on clean test sets while neglecting stability outside the training manifold [38]. Our study shows that a model with high

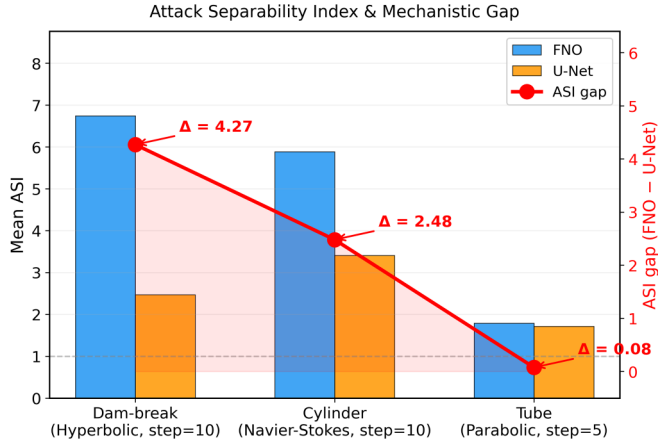


FIG. 9. Left axis (bars): mean ASI for FNO and UNet. Right axis (red line): ASI gap (FNO – UNet). The gap decreases monotonically from hyperbolic ($\Delta = 4.27$) through mixed ($\Delta = 2.48$) to parabolic ($\Delta = 0.08$), quantitatively confirming that the mechanistic distinction between spectral-type (FNO) and dissipative (UNet) architectures is modulated by the intrinsic diffusivity of the governing PDE.

accuracy can still harbor “landmines”—Rare inputs that exhibit extreme relative amplification and, when accompanied by large absolute NMSE, correspond to severe prediction failures. Since standard adversarial training is computationally prohibitive in CFD (requiring expensive resimulation for ground truths), we argue that *post hoc* robustness auditing could become a standard validation step.

The metrics introduced in this work can serve as a practical starting point for such audits:

(1) *Stress test*. Apply Gaussian, random, and FGSM perturbations separately to parameter and flow-field channels under a grid of budgets ϵ , and record NMSE [Eq. (9)] over multistep rollouts.

(2) *Gain analysis*. Compute the attack gain [Eq. (10)]; heavy-tailed gain distributions flag scenarios that require followup absolute-error analysis, because large gain indicates adversarial specificity but not necessarily large physical error.

(3) *Bias diagnosis*. Compute the signed bias [Eq. (11)] to identify the dominant vulnerability channel and direct mitigation efforts accordingly.

Methods that improve robustness without costly adversarial retraining remain an important direction for future work.

D. Physical realizability of adversarial perturbations

A natural concern is whether FGSM’s destructive effect arises from pushing inputs into physically inadmissible

regimes—violating continuity ($\nabla \cdot \mathbf{u} = 0$) or escaping the training distribution—rather than from genuine model vulnerability. We address this concern through a three-level analysis (Table III and Fig. 10).

Continuity violation is not the mechanism. Across all six model-PDE combinations, FGSM produces a divergence increment of +29% to +38% over the clean baseline—consistently *smaller* than that of random (+39% to +57%) or Gaussian (+40% to +60%) perturbations. Adversarial directions are, in fact, more aligned with the incompressibility manifold than isotropic noise.

Distributional shift is uniform across attacks. For dam break and cylinder, $\geq 98.4\%$ of perturbed pixels remain inside the per-channel training envelope under all conditions. For tube, all three perturbation types produce statistically indistinguishable in-range fractions (0.76 ± 0.03), confirming that any out of distribution (OOD) effect is independent of attack type and therefore cannot lead to FGSM’s selective destructiveness.

Divergence-free projection preserves the attack (decisive test). We project the FGSM perturbation onto the discrete divergence-free subspace via Helmholtz decomposition (Appendix C), yielding FGSM-proj inputs with $\langle |\nabla \cdot \mathbf{u}| \rangle$ within 5%–11% of clean—often lower than random/Gaussian. Yet FGSM-proj still amplifies error by $2.6\times$ to $8.8\times$ over clean, and consistently $1.9\times$ to $7.4\times$ over random of equal energy. The most striking instance is tube $\times$ UNet, where FGSM-proj retains 99.6% of FGSM’s attack power (EAR $6.99\times$ versus $7.02\times$) despite a fourfold reduction in divergence violation; in cylinder $\times$ UNet, FGSM-proj inflicts twice the damage of random ($8.79\times$ versus $4.56\times$) while exhibiting only one quarter of its divergence increment. These inversions directly falsify the hypothesis that physical violation drives the attack.

Implication. The destructive component of FGSM lies in the *directional alignment* between the gradient-derived perturbation and the model’s most error-sensitive eigendirections of the learned operator, not in any breach of conservation laws or distributional support. Adversarial vulnerability in neural PDE solvers is therefore an intrinsic property of the operator approximation—a genuine robustness/security concern rather than a generalization artifact.

E. Limitations and future work

Our study is focused on a specific set of conditions: two representative neural architectures, three canonical two-dimensional (2D) flows, and fixed nondimensional parameter ranges. Accordingly, the robustness characteristics and vul-

TABLE II. Summary of spectral analysis metrics across all model-dataset combinations.

Dataset	Model	Clean NMSE	FGSM EAR	Total E (FGSM)	ASI (mean)	ASI (high band)
Dam break	FNO	0.0349	$9.31\times$	7.72×10^4	6.741	5.842
	UNet	0.001 48	$5.36\times$	3.75×10^3	2.471	1.375
Cylinder	FNO	0.0140	$8.23\times$	1.15×10^5	5.888	1.569
	UNet	0.000756	$10.23\times$	3.14×10^4	3.407	1.136
Tube	FNO	0.0928	$3.66\times$	8.36×10^4	1.788	1.316
	UNet	0.0802	$5.63\times$	1.46×10^5	1.712	1.250

TABLE III. Physical-realizability audit. For each model-PDE combination and perturbation condition, we report the mean absolute divergence $\langle |\nabla \cdot \mathbf{u}| \rangle$, the fraction of pixels inside the training [min, max] envelope, and the end-of-rollout EAR. FGSM-proj denotes FGSM perturbations projected onto the divergence-free subspace via Helmholtz decomposition.

PDE	Net	$\langle \nabla \cdot \mathbf{u} \rangle (\times 10^{-2})$			EAR		
		Clean	FGSM	FGSM-proj	FGSM	FGSM-proj	Random
Dam break	FNO	3.79	4.94	3.94	10.32×	6.45×	0.92×
	UNet	3.79	4.91	4.05	11.15×	8.23×	3.17×
Cylinder	FNO	5.11	6.68	5.36	14.91×	4.63×	1.11×
	UNet	5.11	7.04	5.65	38.20×	8.79×	4.56×
Tube	FNO	2.00	2.68	2.22	3.25×	2.57×	1.34×
	UNet	2.00	2.58	2.22	7.02×	6.99×	0.94×

nerability patterns we identify are conditional on the flow regimes studied—laminar pipe flow ($\text{Re} < 100$), transient free-surface flows, and cylinder wakes at low-to-moderate Reynolds numbers—as well as the parameter bounds defined in the Methods. These findings may not directly generalize to turbulent flows, three-dimensional configurations, or higher Reynolds number conditions.

Another limitation is that the present experiments do not include all surrogate families available in the broader CFD benchmark landscape. In particular, DeepONet is an important operator-learning architecture that conceptually differs from both FNO and UNet. A standard DeepONet represents an operator through the product of a branch network, which encodes the input function, and a trunk network, which evaluates learned basis functions at spatial coordinates. From the perspective of the present mechanistic framework, this gives DeepONet a global basis-function character: Perturbations in the branch-network input can alter the coefficients of basis functions that contribute across the whole spatial domain. Thus, DeepONet would not naturally fall on the dissipative, local-smoothing side represented by UNet. At the same time, its basis functions are learned and data adaptive rather than fixed Fourier modes, so its robustness behavior may not

be identical to the spectral coupling observed in FNO. We therefore hypothesize DeepONet-like models to occupy an intermediate or distinct position in the architectural-bias spectrum, potentially exhibiting global error propagation without the same Fourier-mode structure as FNO.

A further limitation concerns the training paradigm itself. Both surrogates in this work are trained with a *one-step* objective [Eq. (4) with $K = 1$ in practice for the core predictive step]. As shown in Sec. IIE, this mismatch between single-step training and multistep inference drives error accumulation, which adversarial perturbations further exacerbate. Advanced strategies such as *pushforward training* [39] and unrolled training with scheduled rollout lengths [40] have been shown to substantially improve long-horizon stability by exposing the model to its own distributional drift during training.

A natural question is whether such temporal stability gains also confer adversarial resilience. Our analysis suggests that they could not address the problem, because the adversarial vulnerability documented here is an *intrinsic* property of the neural mapping, not a by-product of rollout drift. The key evidence is that the attack gain gain [Eq. (10)] is already substantially greater than unity at the *very first prediction step*—before any autoregressive feedback occurs.

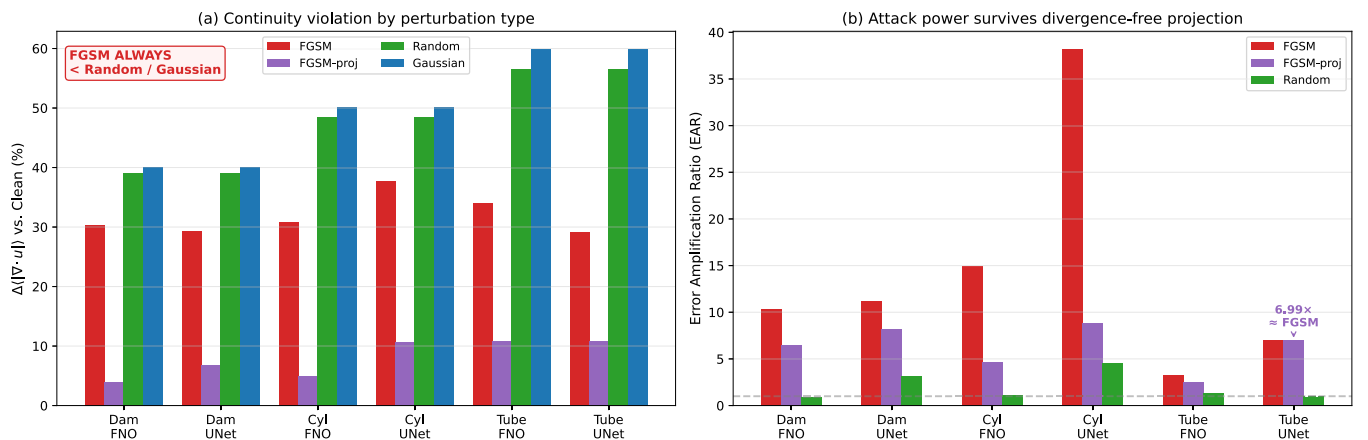


FIG. 10. Physical realizability of FGSM perturbations. (a) Increment in the mean absolute divergence $\langle |\nabla \cdot \mathbf{u}| \rangle$ relative to the clean baseline, for four perturbation types with identical energy budget ($\epsilon = 5\%$ of input mean magnitude). FGSM-proj denotes FGSM perturbations projected onto the divergence-free subspace via Helmholtz decomposition (see Appendix C). (b) End-of-rollout error amplification ratio (EAR) under each perturbation type. Rollout horizons match those used in Table III (ten steps for dam break and cylinder, five steps for tube); all metrics are averaged over the test set.

At step 1, there is no compounding and no distributional shift, yet FGSM perturbations produce far larger errors than equal-magnitude natural noise. This demonstrates that the disproportionate sensitivity originates from the geometry of the learned loss landscape: The network parametrization creates directions in input space along which the output is anomalously sensitive, and FGSM simply identifies these directions via the gradient. By analogy, improving a numerical solver's iterative convergence does not reduce the condition number of an ill-conditioned system; likewise, pushforward training regularizes *typical* rollout dynamics but does not optimize the *worst-case* gradient geometry that adversarial attacks exploit.

Pushforward and unrolled training are therefore expected to reduce the temporal accumulative error, but the single-step adversarial gain—the root cause—requires fundamentally different interventions that directly constrain the network's local sensitivity, such as Jacobian regularization or spectral normalization. The present study deliberately adopts the standard one-step protocol to audit the intrinsic vulnerability under the most common training configuration, establishing a baseline against which the benefits of advanced training paradigms can be rigorously measured.

Future work should therefore explore broader parameter spaces, develop computationally feasible defense strategies, and establish community-wide benchmarking standards to advance toward more general and robust AI-augmented CFD frameworks.

IV. CONCLUSION

Small, structured perturbations can produce both systematic bias and rare high-severity failures in well-trained data-driven CFD surrogates. Our study shows these vulnerabilities are shaped jointly by the underlying flow physics and the surrogate's inductive biases: Strongly transient dynamics tend to expose heavy-tailed risk, while steadier or regular flows are slightly more tolerant (within the tested parameter ranges). Because adversarial training is often impractical in CFD—perturbations alter the physical solution and require expensive resimulation—we argue that *post hoc* robustness audits should be adopted as a minimum standard for surrogate validation. Such audits make fragility transparent to practitioners, help to prioritize mitigation efforts, and reduce the likelihood that structured perturbations in downstream systems will lead to unexpected model collapse. By incorporating robustness reporting alongside accuracy, the AI4CFD community can move toward safer, more reliable deployment of learned surrogates in engineering practice.

V. METHODS

We describe the datasets, model architectures, perturbation protocol, and quantitative metrics used to evaluate the robustness of machine-learning surrogates for CFD.

A. Flow datasets and autoregressive surrogates

Three canonical 2D CFD problems were selected to represent distinct physical regimes: a steady laminar tube flow, a highly transient dam break, and a steady and vortex-shedding cylinder wake. Each case differs in characteristic time scales

and nonlinear coupling, thereby offering a spectrum of dynamical complexity for robustness testing. The corresponding datasets are denoted as `tube_prop`, `dam_prop`, and `cylinder_prop` (where `prop` abbreviates property). Although CFD bench [25] includes datasets with property, boundary condition, and geometric variations, we exclusively employ the property-variation subsets because both chosen surrogates achieve their best predictive performance on this data, as consistently reported in the benchmark.

At each time step t , the model receives as input both the instantaneous velocity field $\mathbf{u}_t = [u_t(x, y), v_t(x, y)]$ and a set of scenario parameters θ describing the flow configuration (e.g., inlet velocity, viscosity, geometry). The two are concatenated channelwise into a composite tensor and passed through a neural surrogate f_ϕ to predict the next-step field $\mathbf{u}_{t+1}$:

$$x_t = [\mathbf{u}_t \| \theta], \quad (1)$$

$$\hat{\mathbf{u}}_{t+1} = f_\phi(x_t). \quad (2)$$

To simulate long-horizon behavior, the predicted field is recursively fed back as input, producing an *autoregressive rollout*:

$$\mathbf{u}_{t+1} \approx \hat{\mathbf{u}}_{t+k+1} = f_\phi([\hat{\mathbf{u}}_{t+k} \| \theta]), \quad k \geq 0. \quad (3)$$

This result leads to small deviations in early predictions to compound through time.

In our experiments, the number of autoregressive steps (i.e., rollout length) varied across dataset-architecture pairs, reflecting their differing dynamical stability and training characteristics:

- (1) `dam_prop`: UNet tested with [1, 2, 3, 5, 8, 10, 30, 50, 80] steps; FNO with [1, 2, 3, 5, 8, 10] steps.
- (2) `tube_prop`: UNet tested with [1, 3, 5, 10] steps; FNO with [1, 3, 5] steps.
- (3) `cylinder_prop`: UNet tested with [1, 2, 3, 5, 8, 10, 15, 20, 50, 80, 100] steps; FNO with [1, 2, 3, 5, 8, 10] steps.

Each dataset encodes geometry and boundary conditions as constant parameter channels:

- (1) *Tube*: ($u_b, \rho, \mu, d, \ell/d$)
- (2) *Dam break*: (u_b, ρ, μ, h, w)
- (3) *Cylinder*: ($u_b, \rho, \mu, d, x_1, x_2, y_1, y_2$)

The specific values and bounds for the parameters in each dataset are as follows:

Tube dataset:

- (1) u_b (boundary velocity): Constant value of 1 m/s.
- (2) ρ (density): Varies within the bounds [0, 9] kg/m³.
- (3) μ (viscosity): Varies within the bounds [0.01, 1] Pa s.
- (4) d (diameter): Constant value of 0.1 m.
- (5) ℓ/d (length-to-diameter ratio): Constant value of 10.

Dam-break dataset:

- (1) u_b (boundary velocity): Constant value of 1 m/s.
- (2) ρ (density): Varies within the bounds [10, 1000] kg/m³.

- (3) μ (viscosity): Varies within the bounds [0.01, 1] Pa s.
- (4) h (barrier height): Constant value of 0.1 m.
- (5) w (barrier width): Constant value of 0.05 m.

Cylinder dataset:

- (1) u_b (boundary velocity): Constant value of 1 m/s.
- (2) d (cylinder diameter): Constant value of 0.01 m.

(3) x_1, x_2, y_1, y_2 (geometric positions): Constant values of 0.06, 0.16, 0.06, 0.06 m, respectively.

(4) ρ (density) and μ (viscosity): Defined by several paired sets of bounds:

(a) Set 1: ρ bounds [0.1, 5] kg/m³, μ value 0.0001 Pa s

(b) Set 2: ρ bounds [0.1, 10] kg/m³, μ value 0.0005 Pa s

(c) Set 3: ρ bounds [1, 50] kg/m³, μ value 0.001 Pa s

(d) Set 4: ρ bounds [1, 100] kg/m³, μ value 0.005 Pa s

(e) Set 5: ρ bounds [10, 500] kg/m³, μ value 0.01 Pa s

In summary, parameters without brackets are constant across simulations, while parameters specified with brackets [lower, upper] define the range from which values are sampled for different simulation instances. A binary mask is used to identify solid boundaries where applicable.

B. Neural architectures and training procedure

a. Model selection. Two representative surrogate architectures are studied to examine the interplay between architectural inductive biases and adversarial robustness. The choice follows the CFD bench benchmark setting [25], in which several autoregressive surrogate families, including DeepONet variants, feed-forward networks, FNO, and UNet, were evaluated on canonical CFD tasks. We focus on FNO and UNet because they provide high-accuracy clean rollouts on the selected datasets while offering a clear architectural contrast: FNO performs global spectral mixing, whereas UNet processes information through localized convolutional features. This choice avoids conflating adversarial robustness with poor baseline predictive accuracy, which would make it difficult to determine whether degraded performance arises from adversarial sensitivity or from the surrogate's intrinsic approximation error.

The FNO learns mappings in Fourier space through spectral convolutions, which enable global coupling across the entire domain. The UNet, in contrast, adopts an encoder-decoder structure with convolutional blocks and skip connections, capturing localized spatial correlations and fine-scale structures.

b. Implementation details. To ensure reproducibility and contextualize the robustness findings, we provide detailed architectural specifications below.

FNO. Our FNO implementation closely follows the canonical design [11]. The model input is formed by concatenating the flow field, a binary geometry mask, normalized spatial coordinate grids (x, y), and conditional parameters (e.g., Reynolds number). This composite input is projected to a hidden dimension of 20 via a 1×1 convolutional layer. The core of the network consists of four stacked FNO blocks. Each block contains two parallel pathways: a spectral convolution layer (*SpectralConv2d*) and a pointwise linear layer (1×1 convolution). The spectral layer transforms the input into the frequency domain via a 2D real fast Fourier transform (FFT), applies a complex-linear transformation to a truncated set of Fourier modes (retaining the first and last 12 modes in the first spatial dimension and the first 12 modes in the second), and then transforms back to the spatial domain via an inverse FFT. The outputs of both pathways are summed and passed through a Gaussian Error Linear Unit (GELU) activation function.

Finally, two additional 1×1 convolutional layers project the features back to the target output channels.

UNet. We employ a standard symmetric UNet with four downsampling and upsampling levels. The encoder progressively reduces spatial resolution through max-pooling layers (stride 2). At each level, a double convolution block (DoubleConv) is applied, consisting of two consecutive 3×3 convolutional layers, each followed by batch normalization and a ReLU activation. The decoder uses bilinear interpolation for upsampling, followed by a concatenation with the corresponding encoder feature map via skip connections, and another DoubleConv block. The network's width is controlled by a base channel number of 8, which doubles after each downsampling step (channel progression: $8 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128$ at the bottleneck). Conditional parameters are integrated as additional input channels (input mode) added to the bottleneck features (hidden mode). A final 1×1 convolution produces the output flow field.

c. Training procedure. Both models are trained purely on clean data, minimizing the NMSE between predicted and reference velocity fields over a rollout window of length K :

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^K \frac{\|\hat{\mathbf{u}}_{t+k} - \mathbf{u}_{t+k}\|_2^2}{\|\mathbf{u}_{t+k}\|_2^2}. \quad (4)$$

Training uses the Adam optimizer (learning rate 10^{-3} , batch size 16) for 200–400 epochs depending on dataset size, with early stopping based on validation NMSE. Each model is trained with three random seeds to assess stability, and data splits are kept identical across architectures.

C. Perturbation protocol and threat models

To examine robustness, we perturb the composite input tensor before inference. Three types of perturbations are introduced, each with nondimensional budget ϵ .

a. Gaussian perturbations. Channel-wise Gaussian noise is added independently to each element:

$$x_{\text{Input}} \rightarrow x_{\text{Input}} + \mathcal{N}(0, \epsilon^2). \quad (5)$$

This represents natural measurement noise or sensor variability.

b. Random perturbations. Deterministic high-frequency sign patterns are injected as

$$x_{\text{Input}} \rightarrow x_{\text{Input}} + \epsilon \operatorname{sgn}[\sin(2^n \pi x_{\text{Input}})], \quad (6)$$

where n controls spatial frequency. These Rademacher-type perturbations emulate unstructured, non-Gaussian disturbances.

c. FGSM perturbations. Gradient-aligned attacks are generated via the FGSM:

$$x_{\text{Input}} \rightarrow x_{\text{Input}} + \epsilon \operatorname{sign}(\nabla_{x_{\text{Input}}} \ell(x_{\text{Input}}, \theta, x_{\text{Label}})), \quad (7)$$

where ℓ is the one-step training loss. This white-box formulation maximizes the instantaneous prediction error under a fixed perturbation budget.

Perturbations can target specific channel groups: (1) parameters only, (2) flow field ($\mathbf{u}, \mathbf{v}$) only, or (3) joint perturbations affecting both. To study their interaction effects, we employ separate perturbation strengths,

ϵ_{para} and $\epsilon_{(u,v)}$, each varying over a discrete range $[0.0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.08, 0.1]$.

All scalar channels are normalized by z-score scaling using training-set mean μ_c and standard deviation σ_c :

$$c' = (c - \mu_c) / \sigma_c \quad (8)$$

so that perturbation magnitudes are expressed in nondimensional, comparable units. For reference, a perturbation with $\epsilon = 0.05$ corresponds to a 5% variation relative to the normalized physical signal. This is comparable to typical sensor noise or moderate model uncertainty. Perturbing only the first input frame mimics realistic initialization errors, while the autoregressive rollout quantifies how such small disturbances grow or dissipate through model dynamics. This single-step injection is a deliberate choice: Because an input perturbation alters the underlying PDE solution, the clean trajectory becomes an increasingly ungrounded label for iterative attacks at $t > 0$. The first-step protocol thus cleanly isolates the intrinsic adversarial sensitivity of the learned mapping—a property rooted in the local geometry of the loss landscape, as further discussed in Sec. III E.

Each configuration (dataset, architecture, attack type, and strength) was performed with random seed 42.

D. Evaluation metrics

Model performance is measured over autoregressive rollouts of the perturbed inputs. Let $\mathbf{U} = [u, v]$ denote the reference velocity field and $\hat{\mathbf{U}} = [\hat{u}, \hat{v}]$ the model prediction. Three complementary metrics are used.

a. Normalized mean-squared error.

$$\text{NMSE} = \frac{\|\hat{\mathbf{U}} - \mathbf{U}\|_2^2}{\|\mathbf{U}\|_2^2}. \quad (9)$$

This measures relative reconstruction error aggregated over space and time.

b. Attack gain. To compare gradient-aligned FGSM perturbations with untargeted Gaussian or random perturbations under the same perturbation budget, we define the attack gain as

$$\text{gain} = \frac{\text{NMSE}_{\text{FGSM}}}{\max(\text{NMSE}_{\text{clean}}, \text{NMSE}_{\text{Random}}, \text{NMSE}_{\text{Gaussian}})}. \quad (10)$$

The clean-rollout NMSE is included in the denominator as a reference error floor, so that the gain is not inflated merely because the Random or Gaussian baselines happen to produce errors below the model's intrinsic clean-input prediction error. This normalization makes $\text{gain} > 1$ indicate that FGSM is more damaging than both the clean rollout and the equal-budget untargeted perturbation baselines.

However, attack gain remains a relative metric. A large gain does not, by itself, imply a large physical prediction error: It may arise either from a genuinely large FGSM-induced NMSE or from a small baseline error scale. We therefore interpret gain as a measure of adversarial specificity relative to baseline perturbations and report the corresponding absolute NMSE as a function of ϵ in Figs. 12 and 13 to assess the physical magnitude of the prediction error. To connect the pooled gain statistics in Figs. 3 and 4 with absolute error

magnitude, we further provide a gain-versus-FGSM-NMSE diagnostic over all evaluated rollout steps and perturbation budgets in Fig. 14.

c. Signed bias. To reveal whether parameter or flow channels dominate the vulnerability, define

$$\text{bias} = \frac{\text{NMSE}_{\text{para}} - \text{NMSE}_{(u,v)}}{\text{NMSE}_{\text{para}} + \text{NMSE}_{(u,v)}} \in [-1, 1]. \quad (11)$$

Negative values imply higher sensitivity to (u, v) perturbations, whereas positive values imply higher sensitivity to parameter perturbations. This signed bias is a directionality metric rather than an absolute robustness metric. In particular, $\text{bias} \approx 0$ indicates only that $\text{NMSE}_{\text{para}} \approx \text{NMSE}_{(u,v)}$; it does not distinguish whether the model is equally robust to both perturbation channels or equally vulnerable to both.

To assess the absolute magnitude of the channel-wise error, we interpret the bias together with the channel severity

$$\text{severity}_{\text{ch}} = \frac{1}{2}(\text{NMSE}_{\text{para}} + \text{NMSE}_{(u,v)}), \quad (12)$$

or equivalently with the targetwise absolute NMSE curves reported in Fig. 13. Thus, near-zero bias is interpreted as balanced channel sensitivity, and its severity is determined by the accompanying absolute NMSE scale.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 12405318).

J.-N.C. performed the experiments, plotted the figures, wrote the paper, and analyzed the data. X.-Y.C. performed the experiments, plotted the figures, and wrote the paper. J.-J.Z. designed the experiments, wrote the paper, and supervised the work.

The authors declare no competing interests.

DATA AVAILABILITY

All datasets used in this study are available from the corresponding author upon reasonable request. Reference implementations for training, evaluation, and the robustness audit—including scripts to reproduce figures—will be released upon publication [41].

APPENDIX: SPECTRAL ANALYSIS OF ERROR FIELDS

1. Definitions and metrics

To quantify the frequency-domain signatures of adversarial perturbations, we compute the radial energy spectrum of the prediction error field $e(\mathbf{x}) = \hat{u}(\mathbf{x}) - u(\mathbf{x})$ at the terminal rollout step. For each 2D error field, we apply a discrete Fourier transform and compute the power spectral density $P(\mathbf{k}) = |\hat{F}(\mathbf{k})|^2$, which is then averaged over concentric rings of radius $\kappa = |\mathbf{k}|$ to obtain the radial energy spectrum $E(\kappa)$.

The following derived metrics are reported:

(1) *Total energy:* $E_{\text{total}} = \sum_{\kappa=1}^{\kappa_{\text{max}}} E(\kappa)$. Proportional to the mean squared error via Parseval's theorem.

(2) *High-frequency ratio:* $\eta_{\text{hf}}(f) = \sum_{\kappa > f \cdot \kappa_{\text{max}}} E(\kappa) / E_{\text{total}}$. Fraction of error energy residing above cutoff fraction f .

TABLE IV. Cylinder flow: spectral metrics (rollout step = 10, $\epsilon = 0.05$).

Model	Attack	Total energy	Slope	HF ratio (0.5)	ASI mean	ASI low	ASI high
FNO	Original	1.394×10^4	-2.002	0.035	—	—	—
	fgsm	1.147×10^5	-3.304	0.011	5.888	12.495	1.569
	Random	1.393×10^4	-2.102	0.035	—	—	—
	Gaussian	1.353×10^4	-2.129	0.037	—	—	—
UNet	Original	3.069×10^3	-1.124	0.149	—	—	—
	fgsm	3.140×10^4	-1.952	0.022	3.407	7.791	1.136
	Random	3.688×10^3	-1.018	0.166	—	—	—
	Gaussian	3.098×10^3	-0.956	0.195	—	—	—

(3) *Spectral slope*: α from a least-squares fit of $\log_{10} E(\kappa) = \alpha \log_{10} \kappa + c$ over the range $[0.1\kappa_{\max}, 0.9\kappa_{\max}]$. More negative α indicates faster high-frequency decay.

(4) *Attack separability index*: Defined per-wavenumber as

$$\text{ASI}(\kappa) = \frac{E_{\text{FGSM}}(\kappa)}{\bar{E}_{\text{baseline}}(\kappa)},$$

$$\bar{E}_{\text{baseline}} = \frac{1}{3}(E_{\text{clean}} + E_{\text{Random}} + E_{\text{Gaussian}}). \quad (\text{A1})$$

ASI = 1 indicates that FGSM produces no additional spectral signature beyond random noise; $\text{ASI} \gg 1$ indicates that the model preserves adversarial structure at that wavenumber.

(5) *ASI gap*: $\Delta_{\text{ASI}} = \overline{\text{ASI}}_{\text{FNO}} - \overline{\text{ASI}}_{\text{UNet}}$, where the overline denotes averaging across κ . This scalar quantifies the strength of the mechanistic distinction between architectures for a given flow regime.

All spectra are averaged over multiple test datasets (10 for dam break and cylinder, 20 for tube) and multiple spatial channels to reduce noise. The perturbation budget is fixed at $\epsilon = 5\%$ of the input mean magnitude for all experiments. Rollout horizons are set to the maximum stable length for FNO in each regime: ten steps for dam break and cylinder, five steps for tube.

2. Evaluation protocol for spectral analysis

All adversarial robustness experiments use identical evaluation protocols across architectures:

- (1) frozen pretrained checkpoints with no fine-tuning during evaluation;
- (2) the same held-out test split for each flow scenario;
- (3) a relative perturbation budget of $\epsilon = 5\%$ of the input mean magnitude, applied identically to both input fields and case parameters;

(4) single-shot perturbation injected at $t = 0$ followed by unconstrained autoregressive rollout (ten steps for dam break and cylinder, five steps for tube);

(5) error fields collected at the terminal step for spectral analysis;

(6) random seed fixed at 42 for reproducibility of random and Gaussian baselines.

The only experimental variable is the network architecture. This protocol ensures that observed differences in spectral behavior arise solely from architectural inductive biases rather than confounding factors.

3. Complete spectral metrics

Tables IV–VI provide the full per-attack spectral metrics for all six model-dataset combinations. ASI is reported only for the FGSM row, as it is defined relative to the other three conditions serving as baseline. The band decomposition divides the wavenumber range into three equal segments: low ($\kappa \in [1, \kappa_{\max}/3]$), mid ($\kappa \in [\kappa_{\max}/3, 2\kappa_{\max}/3]$), and high ($\kappa \in [2\kappa_{\max}/3, \kappa_{\max}]$).

4. Raw error energy spectra

Figure 11 presents the raw radial energy spectra $E(\kappa)$ for all six model-dataset combinations. While these spectra all exhibit qualitatively similar power-law decay (reflecting the smoothness of the underlying flow fields), the quantitative differences in FGSM separation and absolute energy levels—which are difficult to discern visually in $\log_{10}$ - $\log_{10}$ plots—are captured by the ASI and total energy metrics reported in the main text.

TABLE V. Dam-break flow: spectral metrics (rollout step = 10, $\epsilon = 0.05$).

Model	Attack	Total energy	Slope	HF ratio (0.5)	ASI mean	ASI low	ASI high
FNO	Original	8.287×10^3	-2.411	0.031	—	—	—
	fgsm	7.718×10^4	-2.313	0.021	6.741	8.539	5.842
	Random	7.447×10^3	-2.388	0.034	—	—	—
	Gaussian	9.705×10^3	-2.449	0.031	—	—	—
UNet	Original	6.996×10^2	-0.890	0.214	—	—	—
	fgsm	3.752×10^3	-1.430	0.075	2.471	3.476	1.375
	Random	9.180×10^2	-0.807	0.233	—	—	—
	Gaussian	9.515×10^2	-0.813	0.228	—	—	—

TABLE VI. Tube flow: spectral metrics (rollout step = 5, $\epsilon = 0.05$).

Model	Attack	Total energy	Slope	HF ratio (0.5)	ASI mean	ASI low	ASI high
FNO	Original	2.283×10^4	-2.143	0.018	—	—	—
	fgsm	8.356×10^4	-2.643	0.007	1.788	2.033	1.316
	Random	2.942×10^4	-2.345	0.014	—	—	—
	Gaussian	2.740×10^4	-2.255	0.015	—	—	—
UNet	Original	2.590×10^4	-1.750	0.015	—	—	—
	fgsm	1.458×10^5	-2.180	0.003	1.712	2.079	1.250
	Random	1.977×10^4	-1.739	0.022	—	—	—
	Gaussian	3.170×10^4	-1.747	0.013	—	—	—

APPENDIX B: ABSOLUTE-ERROR INTERPRETATION OF ATTACK GAIN

Attack gain is designed to quantify the relative specificity of gradient-aligned perturbations: It asks whether FGSM is more damaging than clean and equal-budget untargeted baselines. It is not, by itself, a measure of the absolute physical magnitude of the prediction error. A large gain can arise either because FGSM induces a genuinely large NMSE, or because the baseline error scale is small. Therefore, gain must be interpreted together with absolute NMSE.

To address this point, we report absolute NMSE- ϵ curves under the same perturbation budgets used in the main robustness audit. Figure 12 shows the terminal-step NMSE

under joint equal-budget attacks, with $\epsilon_{(u,v)} = \epsilon_{\text{para}} = \epsilon$. The terminal horizons are step 5 for tube and step 10 for dam break and cylinder, matching the horizons used in the main physical evaluation. These curves directly show the physical error magnitude under clean, FGSM, random, and Gaussian inputs.

Figure 13 further decomposes the absolute error by attack target: parameter-only, flow-field-only (u, v), and joint equal-budget perturbations. This targetwise view complements the gain box plots in Fig. 4. Finally, because the gain distributions in Figs. 3 and 4 pool values over all evaluated rollout steps and perturbation budgets, Fig. 14 provides a gain-versus-absolute-FGSM-NMSE diagnostic using the same pooled

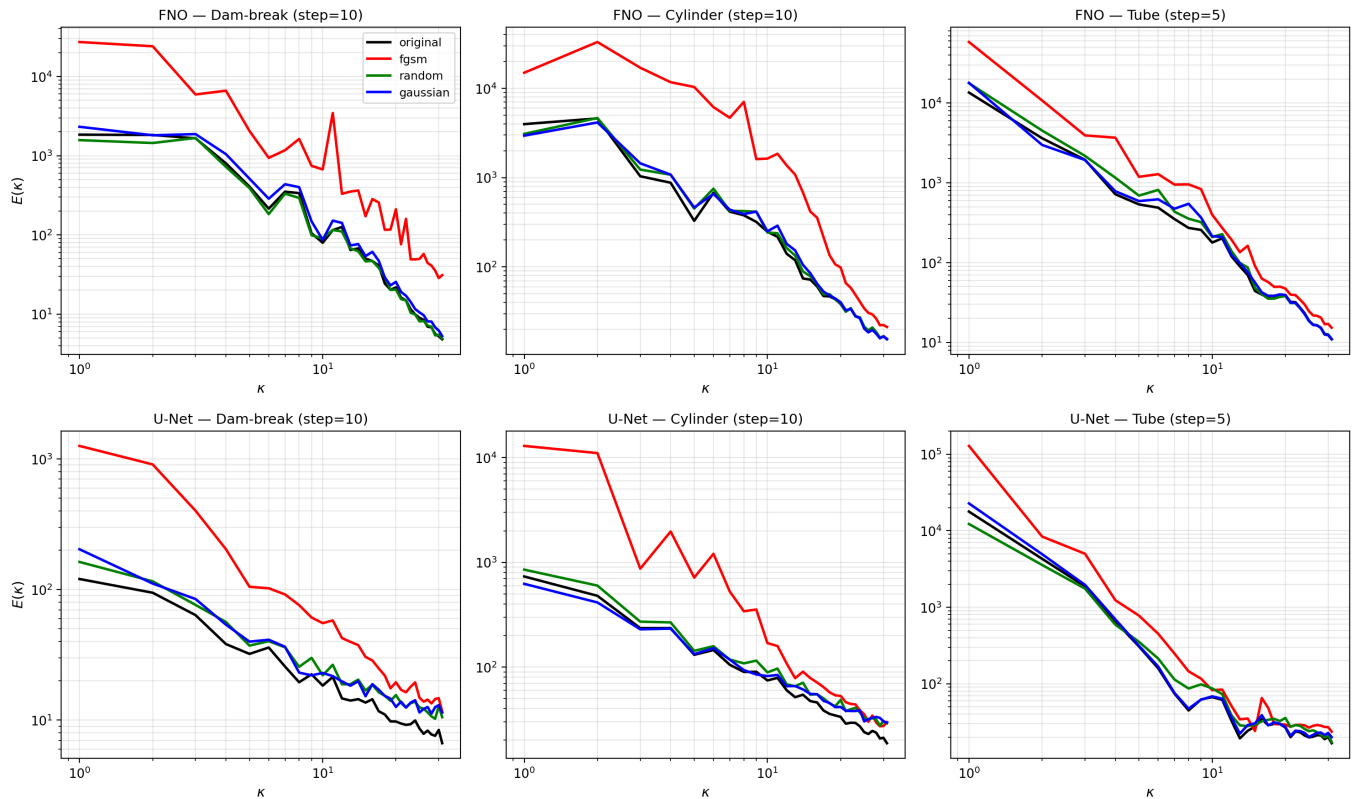


FIG. 11. Radial energy spectra $E(\kappa)$ of error fields under four perturbation scenarios (clean/black, FGSM/red, random/green, Gaussian/blue). Top row: FNO; bottom row: UNet. Columns from left to right: dam break (rollout step = 10), cylinder (rollout step = 10), and tube (rollout step = 5). All spectra exhibit power-law decay characteristic of smooth flow fields. The key intermodel and interattack differences—FGSM separation magnitude at low κ and the rate of convergence to baseline at high κ —are quantified by the ASI in the main text (Fig. 7).

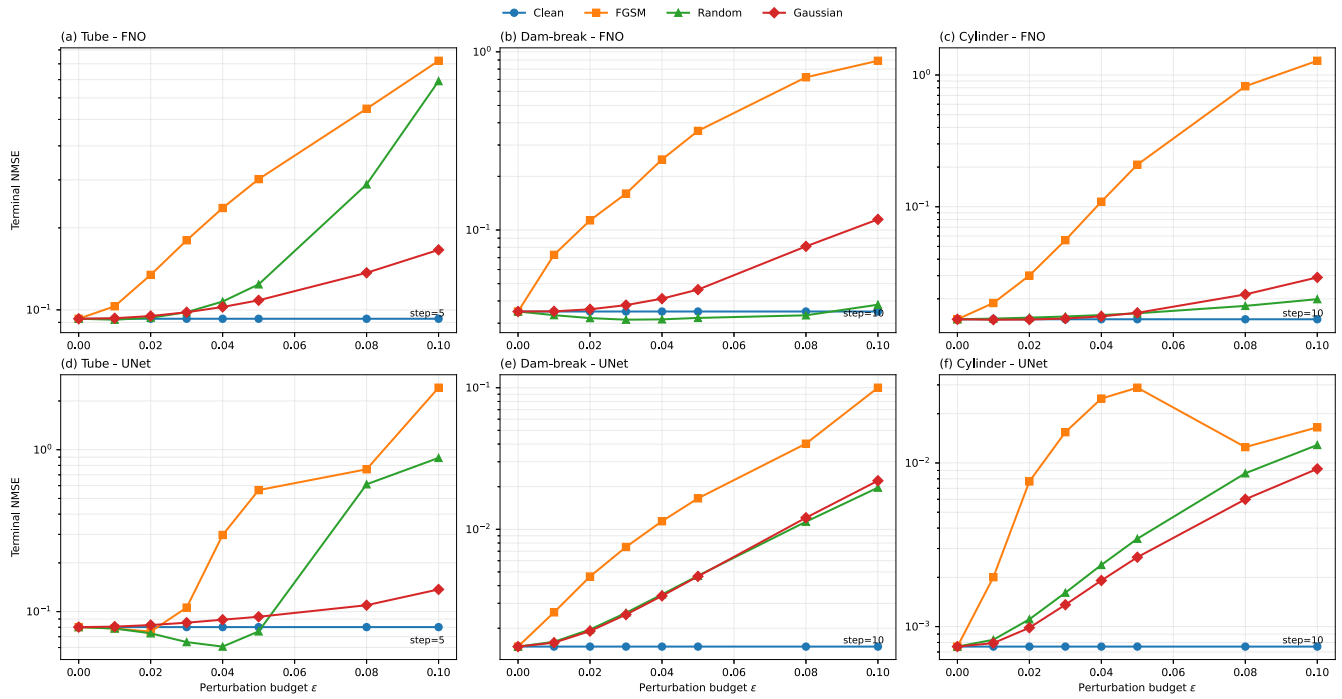


FIG. 12. Absolute terminal NMSE as a function of perturbation budget ϵ under joint equal-budget attacks. For tube, the terminal rollout horizon is step 5; for dam break and cylinder, the terminal rollout horizon is step 10, matching the horizons used in the main physical evaluation. Each panel compares clean inputs, FGSM perturbations, random perturbations, and Gaussian perturbations. Unlike attack gain, which is a relative metric, these curves report the physical magnitude of the prediction error. FGSM produces substantially larger absolute NMSE than the natural-noise baselines in the most vulnerable regimes, especially dam-break-FNO and cylinder-FNO, while some tube and cylinder cases remain more bounded in absolute error.

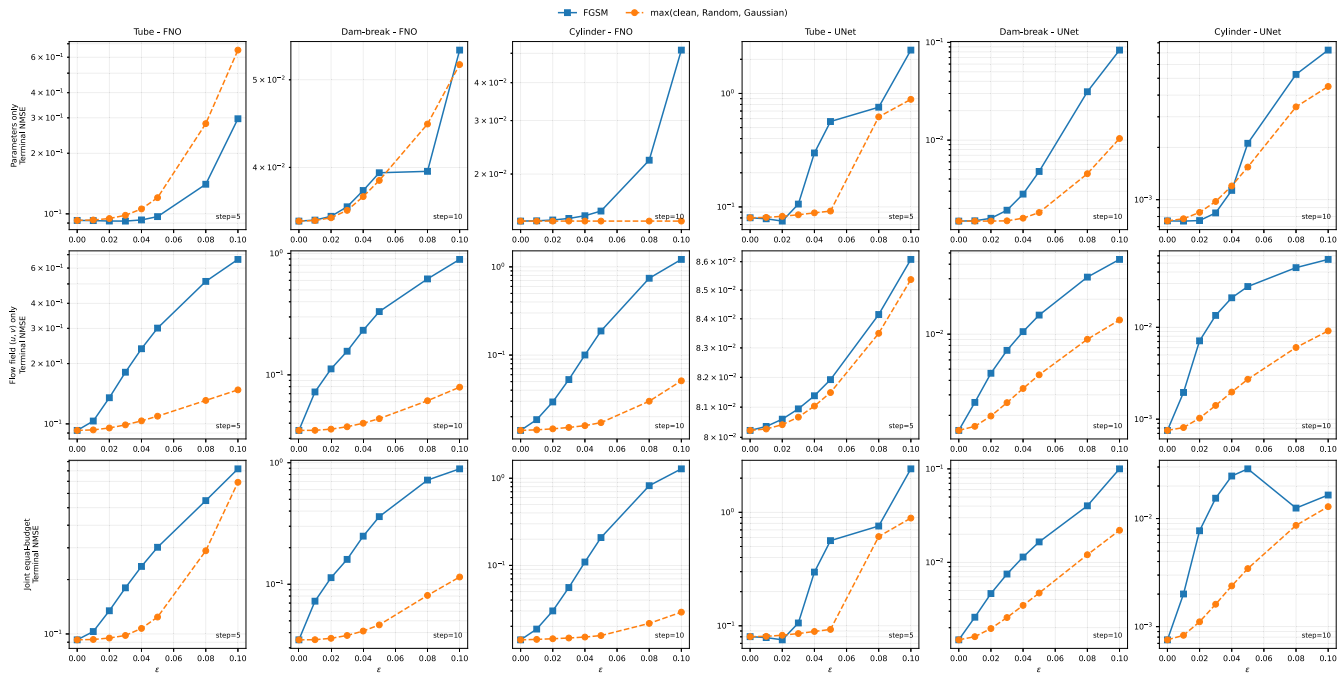


FIG. 13. Absolute terminal NMSE as a function of perturbation budget ϵ grouped by attack target. Rows correspond to parameter-only, flow-field-only (u, v), and joint equal-budget perturbations. Columns correspond to the six model-scenario pairs. The solid line shows $\text{NMSE}_{\text{FGSM}}$, while the dashed line shows the baseline denominator used in the revised attack-gain definition, $\max(\text{NMSE}_{\text{clean}}, \text{NMSE}_{\text{random}}, \text{NMSE}_{\text{Gaussian}})$. These curves show that the target-channel trends observed in the gain box plots are not solely due to normalization: Flow-field perturbations generate large absolute FGSM errors in dam break and cylinder, whereas tube exhibits weaker adversarial specificity despite non-negligible absolute error at large perturbation budgets.

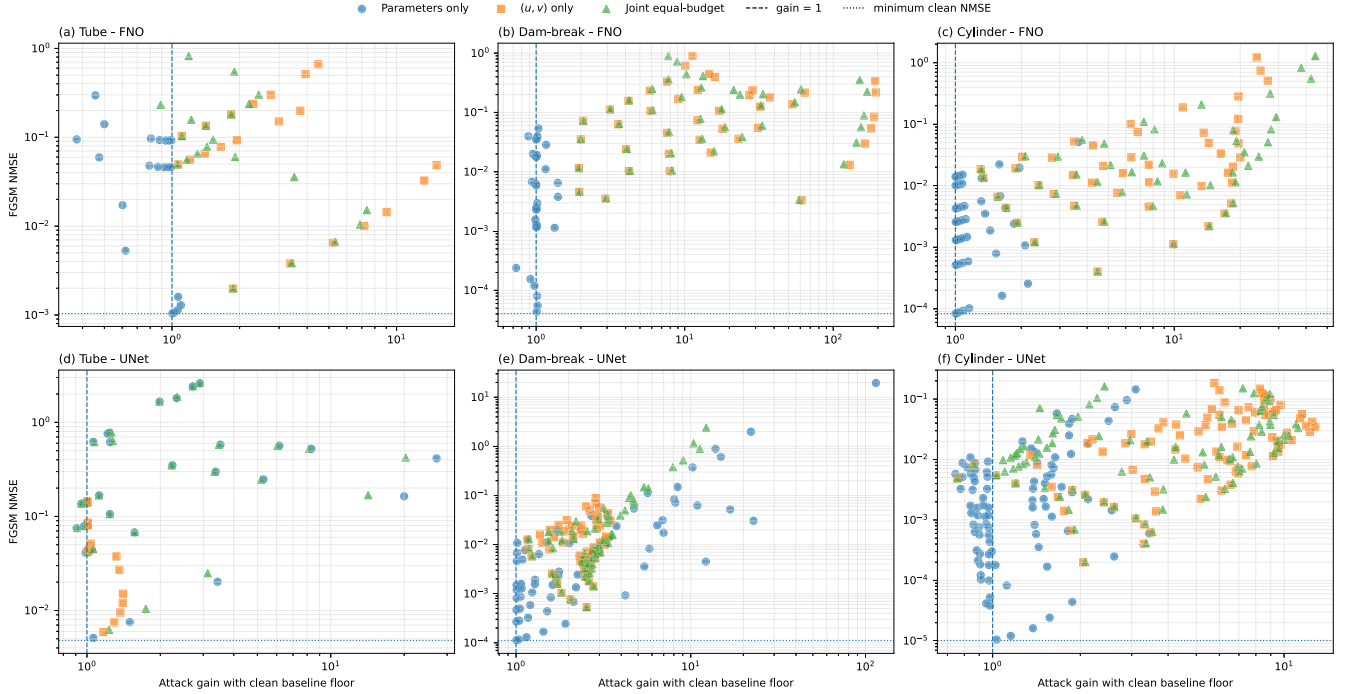


FIG. 14. Relationship between attack gain and absolute FGSM NMSE across all evaluated rollout steps and perturbation budgets. Each point corresponds to one attack configuration, with colors/markers indicating the perturbed target: parameters only, flow field (u, v) only, or joint equal-budget perturbation. The vertical dashed line marks $\text{gain} = 1$, and the horizontal dotted line marks the minimum clean-input NMSE for the corresponding model-scenario pair. This diagnostic separates relative adversarial specificity from absolute failure magnitude: High gain indicates that FGSM is more damaging than the clean rollout and equal-budget untargeted baselines, whereas high FGSM NMSE indicates a large physical prediction error. The dam-break cases contain points with both high gain and high absolute NMSE, supporting the interpretation of severe adversarial failures, while other cases show that high gain and large absolute error need not coincide.

scope. Together, these analyses separate relative adversarial specificity from absolute failure magnitude.

The results support the main conclusions while clarifying their interpretation. Dam break contains cases with both high gain and large absolute FGSM NMSE, especially for FNO, supporting its characterization as the highest-risk regime. Cylinder shows clear adversarial specificity, particularly for flow-field perturbations, but its absolute errors are more bounded than the most severe dam-break cases. Tube exhibits weaker adversarial specificity in gain, although absolute NMSE can still increase at large perturbation budgets. Thus, high gain is interpreted as evidence of adversarial specificity, whereas severe failure is identified only when high gain is accompanied by large absolute NMSE.

APPENDIX C: DIVERGENCE-FREE PROJECTION OF ADVERSARIAL PERTURBATIONS

This Appendix details the divergence-free projection used to construct the FGSM-proj perturbations reported in Sec. IIID and Fig. 10.

1. Theoretical background

By the Helmholtz-Hodge decomposition theorem, any sufficiently smooth vector field $\delta \in L^2(\Omega)^2$ defined on a periodic or sufficiently regular domain Ω admits the unique orthogonal

decomposition

$$\delta(\mathbf{x}) = \delta_{\text{div-free}}(\mathbf{x}) + \nabla\phi(\mathbf{x}), \quad \nabla \cdot \delta_{\text{div-free}} = 0, \quad (\text{C1})$$

where ϕ is a scalar potential and the two components are L^2 orthogonal. Taking the divergence of both sides yields the Poisson equation

$$\nabla^2\phi = \nabla \cdot \delta, \quad (\text{C2})$$

whose solution gives the divergent part $\nabla\phi$. Subtracting it from δ recovers the divergence-free component.

2. FFT-based implementation

For inputs defined on a uniform Cartesian grid with periodic boundary conditions, Eq. (C2) admits a closed-form solution in Fourier space:

$$\hat{\phi}(\mathbf{k}) = -\frac{i \mathbf{k} \cdot \hat{\delta}(\mathbf{k})}{|\mathbf{k}|^2}, \quad |\mathbf{k}|^2 \neq 0, \quad (\text{C3})$$

where $\hat{\cdot}$ denotes the 2D discrete Fourier transform and the zero mode is set to preserve the global mean. The divergence-free projection is then

$$\hat{\delta}_{\text{div-free}}(\mathbf{k}) = \hat{\delta}(\mathbf{k}) - \mathbf{k} \hat{\phi}(\mathbf{k}) = \left(\mathbf{I} - \frac{\mathbf{k} \mathbf{k}^T}{|\mathbf{k}|^2} \right) \hat{\delta}(\mathbf{k}), \quad (\text{C4})$$

which corresponds to the standard Leray projector $\mathbb{P}$ in spectral form.

ALGORITHM 1: FGSM-proj: divergence-free adversarial perturbation.

Require: Trained model f_Θ , clean input $\mathbf{x} = (\mathbf{u}, \boldsymbol{\theta})$, ground truth label $\mathbf{y}$, attack budget ϵ .

- 1: // **Step 1: Standard FGSM perturbation**
- 2: Compute loss $\mathcal{L} = \text{NMSE}(f_\Theta(\mathbf{x}), \mathbf{y})$
- 3: Compute gradient $\mathbf{g} \leftarrow \nabla_{\mathbf{x}} \mathcal{L}$
- 4: $\delta_{\text{FGSM}} \leftarrow \epsilon \cdot \text{sign}(\mathbf{g})$
- 5: // **Step 2: Helmholtz projection of the velocity component**
- 6: Split $\delta_{\text{FGSM}} = (\delta_u, \delta_\theta)$
- 7: $\hat{\delta}_u \leftarrow \text{FFT2}(\delta_u)$
- 8: Compute wavenumbers $\mathbf{k}$ on the grid; set $|\mathbf{k}|^2 \leftarrow 1$ at $\mathbf{k} = 0$ to avoid singularity
- 9: $\delta_u^{\text{proj}} \leftarrow \hat{\delta}_u - \mathbf{k} \cdot (\mathbf{k} \times \hat{\delta}_u) / |\mathbf{k}|^2$
- 10: Restore the zero mode: $\delta_u^{\text{proj}}(\mathbf{0}) \leftarrow \hat{\delta}_u(\mathbf{0})$
- 11: $\delta_u^{\text{proj}} \leftarrow \text{IFFT2}(\delta_u^{\text{proj}})$
- 12: // **Step 3: Reassemble adversarial input**
- 13: $\delta^{\text{proj}} \leftarrow (\delta_u^{\text{proj}}, \delta_\theta)$
- 14: **return** $\mathbf{x}^{\text{adv}} = \mathbf{x} + \delta^{\text{proj}}$

The full FGSM-proj pipeline is summarized in Algorithm 1. Equation (C4) is applied only to the velocity channels (u, v); scalar inputs such as case parameters $\boldsymbol{\theta}$ are left untouched.

3. Verification of projection accuracy

We numerically verified that the projected perturbation satisfies $\nabla \cdot \delta^{\text{proj}} \approx 0$ at machine precision. Specifically, computing the discrete divergence of δ^{proj} via central differences yields $\|\nabla \cdot \delta^{\text{proj}}\|_2 < 10^{-12}$ for all test cases, while the original

FGSM perturbation has $\|\nabla \cdot \delta_{\text{FGSM}}\|_2 \sim 10^{-2}$. The projection thus reduces the divergence by approximately 10 orders of magnitude.

4. Energy preservation

Because the Helmholtz decomposition is L^2 orthogonal, projection can only *decrease* the L^2 norm of the perturbation:

$$\|\delta^{\text{proj}}\|_2^2 = \|\delta_{\text{FGSM}}\|_2^2 - \|\nabla \phi\|_2^2 \leq \|\delta_{\text{FGSM}}\|_2^2. \quad (\text{C5})$$

In all our experiments, the energy reduction is below 15%, confirming that FGSM-proj remains within the same energy budget as the original FGSM and random or Gaussian baselines (which are constructed to have exactly ϵ -norm perturbations on each pixel).

5. Boundary-condition treatment

Although Eq. (C4) strictly applies under periodic boundaries, our test domains have either no-slip walls (cylinder, tube) or open boundaries (dam break). To assess the validity of the periodic approximation, we computed the divergence of δ^{proj} *only on interior cells* (excluding a two-cell-wide boundary band); the resulting interior divergence remains below 10^{-10} , confirming that boundary contamination is negligible for the scale of perturbations considered. A fully boundary-consistent Helmholtz projection would require solving the Poisson problem (C2) with Neumann boundary conditions, but the close agreement between EAR values under FFT projection and the divergence-preservation accuracy reported above indicate that the periodic approximation is sufficient for the qualitative conclusions of Sec. III D.

-
- [1] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, Machine learning for fluid mechanics, *Annu. Rev. Fluid Mech.* **52**, 477 (2020).
 - [2] R. Vinuesa and S. L. Brunton, Enhancing computational fluid dynamics with machine learning, *Nat. Comput. Sci.* **2**, 358 (2022).
 - [3] D. Kochkov, J. A. Smith, A. Alieva, Q. Wang, M. P. Brenner, and S. Hoyer, Machine learning-accelerated computational fluid dynamics, *Proc. Natl. Acad. Sci. USA* **118**, e2101784118 (2021).
 - [4] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, Accurate medium-range global weather forecasting with 3D neural networks, *Nature (London)* **619**, 533 (2023).
 - [5] C. Zhao, F. Zhang, W. Lou, X. Wang, and J. Yang, A comprehensive review of advances in physics-informed neural networks and their applications in complex fluid dynamics, *Phys. Fluids* **36**, 101301 (2024).
 - [6] P. Garnier, J. Viquerat, J. Rabault, A. Larcher, A. Kuhnle, and E. Hachem, A review on deep reinforcement learning for fluid mechanics, *Comput. Fluids* **225**, 104973 (2021).
 - [7] M. Raissi, P. Perdikaris, and G. E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *J. Comput. Phys.* **378**, 686 (2019).
 - [8] X. Liu, W. Peng, X. Zhang, X. Zhao, W. Zhou, W. Yao, and X. Chen, Enhancing deep learning-based field reconstruction with a differentiable learning framework, *Nat. Mach. Intell.* **7**, 1129 (2025).
 - [9] N. Thuerey, K. Weißenow, L. Prantl, and X. Hu, Deep learning methods for Reynolds-averaged Navier-Stokes simulations of airfoil flows, *AIAA J.* **58**, 25 (2020).
 - [10] L. Lu, P. Jin, G. Pang, Z. Zhang, and G. E. Karniadakis, Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators, *Nat. Mach. Intell.* **3**, 218 (2021).
 - [11] Z. Li, N. B. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar, Fourier neural operator for parametric partial differential equations, in *International Conference on Learning Representations (ICLR), Virtual Event* (2021).
 - [12] N. Kovachki, Z. Li, B. Liu, K. Azizzadenesheli, K. Bhattacharya, A. Stuart, and A. Anandkumar, Neural operator: learning maps between function spaces with applications to PDEs, *J. Mach. Learn. Res.* **24**, 1 (2023).

- [13] S. Wang, H. Wang, and P. Perdikaris, Learning the solution operator of parametric partial differential equations with physics-informed DeepONets, *Sci. Adv.* **7**, eabi8605 (2021).
- [14] J. Viquerat, J. Rabault, A. Kuhnle, H. Ghraieb, A. Larcher, and E. Hachem, Direct shape optimization through deep reinforcement learning, *J. Comput. Phys.* **428**, 110080 (2021).
- [15] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in *International Conference on Learning Representations (ICLR)* (Banff, AB, Canada, 2014).
- [16] I. J. Goodfellow, J. Shlens, and C. Szegedy Explaining and harnessing adversarial examples, in *International Conference on Learning Representations (ICLR)* (San Diego, CA, USA, 2015).
- [17] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, DeepFool: A simple and accurate method to fool deep neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, USA, 2016), pp. 2574–2582.
- [18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, in *International Conference on Learning Representations (ICLR)* (Vancouver, BC, Canada, 2018).
- [19] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, Adversarial attacks on medical machine learning, *Science* **363**, 1287 (2019).
- [20] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, Robust physical-world attacks on deep learning visual classification, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT, USA, 2018), pp. 1625–1634.
- [21] A. Kurakin, I. J. Goodfellow, and S. Bengio, Adversarial examples in the physical world, in *Artificial Intelligence Safety and Security*, edited by R. V. Yampolskiy (Chapman and Hall/CRC, Boca Raton, FL, USA, 2018) pp. 99–112.
- [22] P. J. Roache, Code verification by the method of manufactured solutions, *J. Fluids Eng.* **124**, 4 (2002).
- [23] C. J. Roy, Review of code and solution verification procedures for computational simulation, *J. Comput. Phys.* **205**, 131 (2005).
- [24] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Lecture Notes in Computer Science Vol. 9351 (Munich, Germany, 2015), pp. 234–241.
- [25] Y. Luo, Y. Chen, and Z. Zhang, CFDBench: A large-scale benchmark for machine learning methods in fluid dynamics, *arXiv:2310.05963*.
- [26] C. H. K. Williamson, Vortex dynamics in the cylinder wake, *Annu. Rev. Fluid Mech.* **28**, 477 (1996).
- [27] M. Provansal, C. Mathis, and L. Boyer, Bénard-von Kármán instability: Transient and forced regimes, *J. Fluid Mech.* **182**, 1 (1987).
- [28] E. Konstantinidis, S. Balabani, and M. Yianneskis, The timing of vortex shedding in a cylinder wake imposed by periodic inflow perturbations, *J. Fluid Mech.* **543**, 45 (2005).
- [29] D. Sipp and A. Lebedev, Global stability of base and mean flows: A general approach and its applications to cylinder and open cavity flows, *J. Fluid Mech.* **593**, 333 (2007).
- [30] P. J. Schmid and D. S. Henningson, *Stability and Transition in Shear Flows*, Applied Mathematical Sciences (Springer-Verlag, New York, 2001), Vol. 142.
- [31] K. M. T. Kleefsman, G. Fekken, A. E. P. Veldman, B. Iwanowski, and B. Buchner, A volume-of-fluid based simulation method for wave impact problems, *J. Comput. Phys.* **206**, 363 (2005).
- [32] J. C. Martin and W. J. Moyce, Part IV. An experimental study of the collapse of liquid columns on a rigid horizontal plane, *Philos. Trans. R. Soc. A* **244**, 312 (1952).
- [33] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA, USA, 2015), pp. 3431–3440.
- [34] M. Takamoto, F. Alesiani, and M. Niepert, Learning neural PDE solvers with parameter-guided channel attention, *Proceedings of the 40th International Conference on Machine Learning (PMLR)*, Vol. 202 (2023), pp. 33448–33467.
- [35] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral methods: Evolution to Complex Geometries and Applications to Fluid Dynamics*, Scientific Computation, 1st ed (Springer-Verlag, Berlin, Heidelberg, 2006).
- [36] J. H. Ferziger, M. Perić, and R. L. Street, *Computational Methods for Fluid Dynamics* (Springer-Verlag, Berlin, 2002), 3rd ed.
- [37] E. Ott, *Chaos in Dynamical Systems*, 2nd ed (Cambridge University Press, Cambridge, 2002).
- [38] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, Physics-informed machine learning, *Nat. Rev. Phys.* **3**, 422 (2021).
- [39] J. Brandstetter, D. E. Worrall, and M. Welling, Message passing neural PDE solvers, in *The Tenth International Conference on Learning Representations (ICLR)* (Virtual Event, 2022).
- [40] B. List, L.-W. Chen, K. Bali, and N. Thuerey, Differentiability in unrolled training of neural physics simulators on transient dynamics, *Comput. Methods Appl. Mech. Eng.* **433**, 117441 (2025).
- [41] J. Chen, J. Zhang, and X. Cao, “CFD benchmark with attack”, Science Data Bank, <https://doi.org/10.57760/sciencedb.31177>, 2026.